MASTER EN SCIENCES MATHÉMATIQUES

Description et modélisation de la diminution des ressources trophiques sur la biocénose, en Meuse belge et française

SCHMIT, Carol-Ann

Award date:
2015

Awarding institution:
Universite de Namur

UNIVERSITE DE NAMUR
Faculté des Sciences

Description et modélisation de la diminution des ressources trophiques sur la biocénose, en Meuse belge et française.

Mémoire présenté pour l'obtention
du grade académique de master en « sciences mathématiques »

Carol-Ann SCHMIT

Juin 2015

UNIVERSITE DE NAMUR
Faculté des Sciences

Description et modélisation de la diminution des ressources trophiques sur la biocénose, en Meuse belge et française.

Promoteur: Marcel Remon
Copromoteur: Frédérik De Laender
Encadrant: Adrien Latli

Mémoire présenté pour l'obtention
du grade académique de master en « sciences mathématiques »

Carol-Ann SCHMIT

Juin 2015

Remerciements

Tout d'abord, je tiens à remercier Marcel Rémon pour m'avoir permis de réaliser un mémoire dans un domaine qui m'intéresse énormément : les statistiques. Un grand merci à Frederik De Laender et Adrien Latli d'avoir proposé un sujet aussi passionnant que la biodiversité en milieu aquatique.

Du point de vue mathématique, je remercie Marcel Rémon pour son aide et ses conseils à propos de l'application des méthodes statistiques utilisées, ainsi que Adrien Latli pour m'avoir proposé de nombreuses méthodes qui m'étaient inconnues, m'avoir aidé à les mettre en place et à les interpréter.

Je remercie également André Hardy pour m'avoir fait découvrir les statistiques telles que je les connais maintenant. Ses cours, ainsi que celui de M. Rémon, ont été une base solide pour entamer ce mémoire.

Un grand merci à M. Rémon et Fanny Jacquemart pour leur relecture assidue et les corrections orthographiques bien utiles qu'ils ont apportées.
Je remercie également Cédric Charlot et Aurélie Goffin pour leur aide lors de l'écriture anglophone de mon résumé.

Enfin, j'aimerais remercier les master2 math et mon copain pour m'avoir donné des idées, des coups de pouce et m'avoir soutenu tout au long de ce mémoire et de mes études. Je remercie également ma famille pour leur soutien et sans qui je n'aurais peut-être jamais pu entamer d'études en mathématiques.

Résumé

L'effondrement mondial de la biodiversité est une des problématiques majeures du 20ème siècle, dont les trois causes principales sont la destruction des habitats naturels, la surexploitation des ressources et l'invasion d'espèces exotiques. L'arrivée de bivalves exogènes en Meuse belge et française inquiète les biologistes : ces espèces ont-elles un réel impact sur la diminution de certaines populations autochtones et de divers paramètres physicochimiques de la Meuse ?

L'unité de recherche en biologie environnementale évolutive (URBE) de l'UNamur nous a fourni des bases de données concernant les mesures des paramètres biologiques et physicochimiques sur diverses stations positionnées le long de la Meuse. Après avoir choisi les stations que nous étudions, nous analysons l'impact des bivalves exogènes sur la biocénose de la Meuse. Les diverses analyses effectuées sont une analyse des tendances, des régressions linéaires et une analyse de co-inertie.

Mots clés : biodiversité, bivalves, biocénose, Meuse, statistiques, tendance, régression, classification, co-inertie

Abstract

The Global Biodiversity collapse is one of the major issues of the 20th century. The three main causes of this collapse are the destruction of the habitat, the overexploitation of resources and the devlopment of exotic species. The arrival of exogenous bivalvia in Belgian and French Meuse river worries biologists : do they really have an impact on the reduction of some indigenous population and various physicochemical parameters of the river Meuse ?

The research unit in evolutionary environmental biology (URBE : unité de recherche en biologie environnementale évolutive) of the University of Namur provided us a database on containing biological measures and physicochemical parameters on various stations positionned along the river Meuse. After choosing the stations, we analysed the impact of exogenous bivalvia on the biocenosis of the river Meuse though various analyzes such as trend analysis, linear regressions and analysis of co-inertia.

Keywords : biodiversity, bivalvia, biocenosis, Meuse, statistics, trendtest, régression, classification, co-inertia

Table des matières

Introduction

L'unité de recherche en biologie environnementale évolutive (URBE) de l'UNamur a pour vocation d'étudier les organismes aquatiques et leurs interactions avec l'environnement. C'est en collaboration avec cette unité que ce mémoire est développé, l'objectif étant d'évaluer les corrélations existantes entre l'apparition d'espèces envahissantes, la disponibilité des ressources trophiques* [1] et les peuplements autochtones de la rivière Meuse.

Cette étude est réalisée à partir de différentes données fournies par l'unité URBE, basées sur des relevés historiques existants de l'Agence de l'Eau, l'ONEMA, la DEMNA, l'UNamur et l'ULg. Depuis l'arrivée d'espèces exogènes* de bivalves*, les biologistes observent une chute du taux de chlorophylle a, ainsi que de certaines espèces de poissons. Le but de ce mémoire est de prouver, par diverses méthodes statistiques, qu'une réelle chute de chlorophylle a est observée, ainsi que d'autres paramètres. Il faut ensuite démontrer le lien existant entre l'arrivée des espèces exogènes et la chute des paramètres.

Notons qu'un travail de recherche et de lecture de diverses publications a été réalisé afin de comprendre la problématique à laquelle font face les biologistes, ainsi que pour saisir toutes les subtilités des divers mécanismes liant les différentes variables physiques, biologiques et physico-chimiques d'un milieu aquatique. De plus, même si les méthodes statistiques « classiques »sont applicables en biologie, de nombreuses méthodes spécifiques à ce domaine existent. Certaines méthodes ont pris un certain temps à être maitrisées, et certaines ont du être abandonnées car elles ne convenaient pas à nos données.

Le premier chapitre est consacré à la mise en situation de la problématique rencontrée par les biologistes. Il présente la problématique liée à la présence d'espèces de bivalves exogènes* et leur impact sur la biocénose*. Il met également en place les divers concepts biologiques utiles à la compréhension de notre étude.

Le deuxième chapitre porte sur la description des données disponibles ainsi que sur les problèmes rencontrés avec ces dernières et les solutions envisagées.

Le chapitre 3 porte sur les tests de tendance. Dans un premier temps, le concept de série temporelle est défini, ainsi que celui de non stationnarité de ces séries. Le problème de la présence d'autocorrélations dans les données est abordé et différents tests de tendance sont présentés de façon théorique.

Le quatrième chapitre est une application des concepts du chapitre précédent sur nos jeux de données. Les données sont tout d'abord modifiées pour pouvoir être utilisées, ensuite, deux tests de tendance leurs sont appliqués. Une comparaison des deux tests est réalisée ainsi qu'une analyse des résultats pour la méthode retenue.

Le chapitre suivant présente différentes méthodes de classification et méthodes de choix du nombre de classes. Ensuite un chapitre, théorique lui aussi, est consacré à un bref rappel des grandes lignes de la théorie des régressions linéaires pour ensuite aborder les modèles autorégressifs.

Le chapitre 7 est une application des deux chapitres précédents aux données. Dans un premier temps, une classification est effectuée afin de différencier plusieurs périodes dans nos séries de données. Ces périodes sont relatives à l'apparition des bivalves exogènes étudiés et leur installation en Meuse. Une régression est ensuite effectuée afin de mesurer l'impact de

1. Les mots marqués d'une étoile sont définis en annexe A.

ces espèces exogènes et des autres paramètres physicochimiques et biologiques sur le taux de chlorophylle a.

Enfin, le dernier chapitre aborde la co-inertie. Celle-ci permet de représenter nos données en insistant sur l'évolution temporelle de celles-ci.

Chapitre 1

Mise en situation

L'effondrement mondial de la biodiversité est une des problématiques majeures du 20ème siècle, dont les trois causes principales sont la destruction des habitats naturels, la surexploitation des ressources et l'invasion d'espèces exotiques [1].

Nous étudions l'évolution de la biodiversité de la Meuse des années 1980 à nos jours. La Meuse fait partie des réseaux hydrographiques européens qui sont parmi les écosystèmes les plus touchés suite aux aménagements qu'ils ont subis, ceux-ci facilitant la propagation de nouvelles espèces. D'autre facteurs anthropiques* [1] ont eu un effet négatif sur la biocénose*, comme par exemple la dégradation de la qualité de l'eau.

La plupart des espèces exigeantes ont disparu des grands cours d'eau ou se cantonnent à des zones refuges et seules les espèces les plus tolérantes perdurent dans les milieux les plus dégradés. Les autres espaces ont été colonisés par de nombreuses espèces exogènes* disposant d'un moyen de reproduction très efficace et qui sont très tolérantes aux conditions physicochimiques médiocres et à la dégradation physique de l'habitat [38]. Ce type de reproduction est une stratégie de type « r » : *les espèces misent sur la reproduction avec un fort taux de croissance, pour compenser par le nombre* la fragilité due à la perturbation du milieu [8]. Ces espèces invasives entrent en concurrence directe (alimentaire ou habitationnelle) avec les espèces autochtones et peuvent, dans certains cas, atteindre une telle densité qu'ils perturbent l'écosystème entier en modifiant la chaine alimentaire présente, aussi appelée réseau trophique*.

Au cours du 20ème siècle, plusieurs espèces de bivalves invasifs ont successivement colonisés la majorité des grands cours d'eau européens et américains. La moule zébrée (Dreissena polymorpha), le genre Corbicula et plus récemment la moule quagga (Dreissena rostriformis bugensis) forment des populations très denses qui ont causé de nombreuses perturbations écologiques [42]. Ces bivalves invasifs sont des bivalves filtreurs, cela signifie qu'ils filtrent l'eau afin d'y trouver leur nourriture. Ces sont des filtreurs très efficaces, ils sont ainsi en concurrence directe avec les bivalves autochtones. Par exemple, la moule zébrée peut former des « récifs »très épais et compacts, qui peuvent contenir jusqu'à 200 individus par mètre carré [9]. Des études américaines ont démontré que ces colonies de bivalves étaient si denses, qu'elles ont changé les densités (chute de la chlorophylle a) et la composition spécifique du plancton de l'Hudson et ce malgré la taille importante de ce fleuve [36]. Descy et al. (2003) ont observé des résultats analogues sur la Moselle (affluent français du Rhin) et ont modélisé l'impact de cette filtration [45]. Les résultats de la simulation suggèrent que la filtration des mollusques invasifs influence fortement le potamoplancton*, la turbidité*, la qualité de l'eau et le cycle du carbone, notamment aux endroits où ces bivalves sont les plus abondants. Venus d'Europe de l'est, ils sont arrivés en Europe de l'ouest et Amérique du Nord en se fixant sur la coque des bateaux et colonisant peu à peu de nombreux canaux maritimes. Ils causent de graves problèmes à certains utilisateurs d'eau en obstruant des conduites ou en bloquant des écluses et, comme expliqué précédemment, peuvent supplanter puis éliminer d'autres espèces moins résistantes.

1. Les mots notés d'une étoile sont définis en annexe A.

Les bivalves se nourrissent de phytoplancton. Le phytoplancton dénote l'ensemble des organismes végétaux vivant en suspension dans l'eau [22]. Étant la base de la chaine alimentaire en milieu aquatique peu profond, la consommation importante par les bivalves invasifs du phytoplancton n'aurait pas des conséquences seulement sur ses concurrents trophiques* directs, mais sur l'ensemble de la biocénose des cours d'eau touchés. La figure 1.1 représente la chaine alimentaire en milieu aquatique. Les grands prédateurs mangent les plus petits poissons, qui mangent le zooplancton* (plancton animal), qui mange lui même le phytoplancton. Les déchets organiques (plantes, animaux morts) sont décomposés par des bactéries en matières minérales qui sont assimilées par le phytoplancton lors de la photosynthèse dont la formule est la suivante :

$$\text{eau} + CO_2 + \text{énergie solaire} \Rightarrow \text{matière végétale} + O^2.$$

En plus d'être la base de la chaine alimentaire aquatique, le phytoplancton constitue près de la moitié de la production d'oxygène mondiale [22].

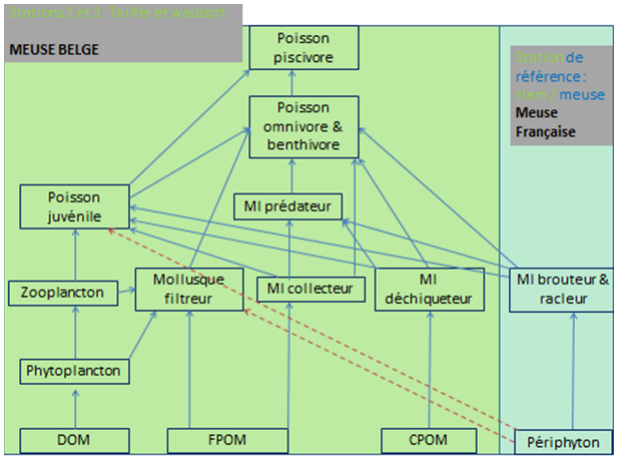


FIGURE 1.1 – Chaine alimentaire en milieu aquatique

Au cours de la dernière décennie, la biomasse phytoplanctonique de la Meuse française et belge est en diminution constante alors que les principaux facteurs abiotiques restent relativement stables (Falisse, 2011). Pigneur et al. (2013) ont également montré une très forte corrélation entre la diminution du phytoplancton sur la Meuse et les densités de mollusques invasifs. De plus, une chute drastique des populations de nombreuses espèces typiques de poissons de la Meuse a été constatée en Meuse mitoyenne [47] alors que trois espèces de bivalves invasifs ont été recensées.

C'est dans ce contexte que nous essayons de montrer une chute du phytoplancton en Meuse belge et française ainsi qu'une corrélation avec l'apparition des espèces invasives de bivalves.

Chapitre 2

Description des données

En une trentaine d'années, la biocénose* de la Meuse a profondément évolué. De nombreux taxons* ont disparu tandis que des espèces exogènes* se sont implantées, parfois massivement, dans un milieu perturbé par une série d'aménagements anthropiques*. L'objectif est de faire ressortir de la masse de données historiques disponibles, les conséquences de « l'invasion »de ces espèces exogènes*, à l'aide d'outils statistiques. Pour cela, nous disposons de deux types de données : les paramètres biologiques (population de macroinvertébrés et population piscicole) et celles concernant les paramètres physicochimiques de l'eau. Les données sont basées sur des relevés historiques existants : Agence de l'Eau, ONEMA, DEMNA, UNamur, ULg. Ces données ont été mesurées des années 1980 à nos jours aux différentes stations se situant le long de la Meuse.

1 Les paramètres physicochimiques

Nous avons extrait des bases de données existantes les paramètres physicochimiques qui influencent le plus la biocénose*. Ces données présentent le minimum, maximum, la moyenne et 90% des variables suivantes :

- le débit Q $[m^3/s]$ du cours d'eau
- la température t (degrés Celsius) de l'eau
- la chlorophylle a $Chla$ $[mg/L]$
- la concentration de matières en suspension mes $[mg/L]$
- le taux d'ammonium NH_4^+ $[mg/L]$
- le taux de nitrates NO_3^- $[mg/L]$
- la concentration en phosphate PO_4^{3-} $[mg/L]$
- la concentration en phosphore $Ptot$ $[mg/L]$
- la concentration en oxygène O_2 $[mg/L]$
- le potentiel hydrogène pH $[mg/L]$

Le rapport entre ces paramètres et notre étude se situe d'une part dans le cycle de l'azote : le phytoplancton utilise, pour se développer, les nitrates NO_3^-, l'ammonium NH_4^+ et le phosphore $Ptot$. Le phosphore se retrouve dans l'eau suite à l'érosion de roches contenant des minéraux phosphatés. D'autre part la chlorophylle a $Chla$ est le principal pigment photosynthétique du règne végétal. La mesure de sa concentration donne donc une bonne estimation de la biomasse* de phytoplancton présente.

2 Les macroinvertébrés

Les invertébrés sont traditionnellement étudiés car ils constituent de bons indicateurs de la qualité globale de l'écosystème aquatique et sont facilement exploitables. Ils forment également un maillon essentiel du réseau trophique entre producteur primaire* et consommateur secondaire*. Les macroinvertébrés sont triés selon la classification scientifique des espèces, ou

« classification biologique », représentée par une pyramide sur la figure 2.1. Les macroinvertébrés font tous partie du règne animal, ce dernier n'est donc pas mentionné dans nos donnée. Les classes mentionnées sont : l'embranchement, la classe, la famille, le genre et parfois l'ordre.

Les macroinvertébrés sont exprimés en nombre d'individus pour les stations françaises et en abondance pour les stations belges. L'abondance est comprise entre 0 et 6 et correspond à un intervalle d'individus dans lequel se situe l'espèce associée. Notons que si une espèce de macroinvertébré n'est pas pêchée, elle sera mentionnée comme ayant 0 individu, cela ne veut pour autant pas dire qu'il n'y en a aucuns.

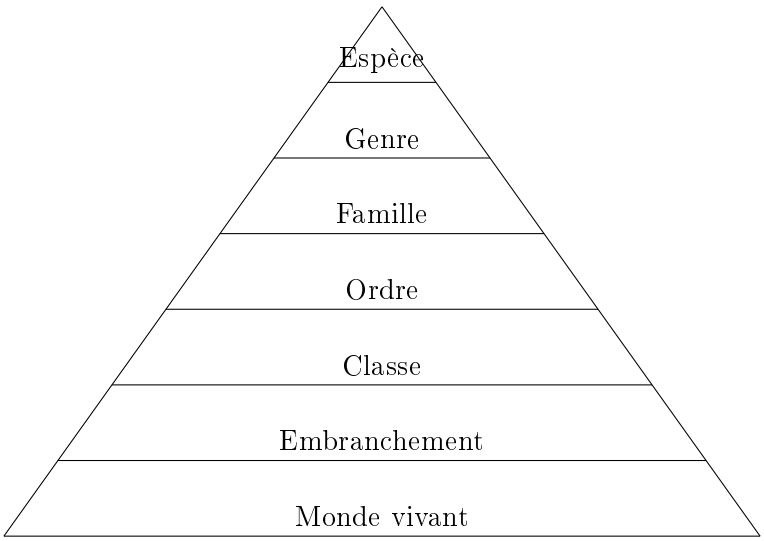


FIGURE 2.1 – Pyramide de la classification biologique

Les données sont uniformisées pour être toutes exprimées en nombre d'individus (et non en abondance). Si une classe d'abondance contient (par définition) entre n et m individus, le nombre d'individus est fixé à $\frac{m+n}{2}$ individus. De cette manière, les stations belges et françaises sont directement comparables.

Les méthodes de recensement

Dans les tableaux de données, la méthode de recensement utilisée est indiquée pour chaque station où les mesures ont été prises. Il en existe plusieurs, présentées ci-dessous :

- IBGA (indice biologique global adapté) : Les zones de prélèvement sont choisies, il existe ensuite plusieurs manières de prélever les macroinvertébrés :
  - les substrats artificiels : des sacs remplis de pierres, similaires à celui présenté en figure 2.2, sont suspendus par une corde dans l'eau, pour ensuite être remontés.

FIGURE 2.2 – Matériel de substrat artificiel

  - le grappin : méthode utilisée pour faire des prélèvements de fond si on peut accéder à la verticale (via un pont ou une écluse) au milieu du cours d'eau. Un grappin similaire à celui présenté en figure 2.3 est alors utilisé.

FIGURE 2.3 – Grappin utilisé pour le recensement des macroinvertébrés

  - échantillonnage sur berge : si la collecte manuelle de pierres de taille moyenne est impossible, cette méthode consiste à passer un filet racloir sur les murs en pierre ou en béton des bords du cours d'eau.
  - drague : méthode utilisée si l'accès au milieu du cours d'eau est impossible sans bateau. Celui-ci remorque une drague (filet lourd de minimum 25 kg) face au courant comme présenté en figure 2.4.

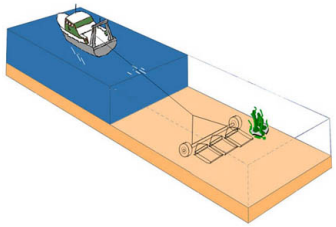


FIGURE 2.4 – Schéma de la méthode de drague

- I2M2 (RCS) : méthode de recensement uniquement utilisée dans les rivières praticables à pied. Un cadre métallique de même type que celui de la figure 2.5 est mis dans le substrat (sol formant le fond de la rivière) afin de réaliser 12 échantillons.
- IBGN (indice biologique global normalisé) : méthode équivalente à I2M2 mis à part que 8 échantillons sont réalisés au lieu de 12.

FIGURE 2.5 – Cadre métallique utilisé pour les méthodes I2M2 et IBGN

3 Travail sur les données

3.1 Choix des stations étudiées

Les stations sont les lieux où les mesures des paramètres biologiques et physicochimiques ont été réalisées. Notre but étant d'étudier la Meuse belge et française, nous avons choisi, de façon stratégique, différentes stations afin de couvrir au mieux la Meuse. Nous avons aussi essayé d'avoir des stations dont les méthodes de recensement soient au maximum les mêmes, ou proches. C'est pourquoi nous avons choisi les stations de Saint-Mihiel, Inor et Ham-sur-Meuse pour la France et celles de Hastière et Liège pour la Belgique, présentées sur la carte en figure 2.6.

Les stations de mesure des macroinvertébrés et des paramètres physicochimiques ne sont pas exactement les mêmes. Ainsi les mesures de macroinvertébrés seront faites à Hastière tandis que celles concernant les paramètres physicochimiques à Tailfer, de même pour les mesures des paramètres physicochimiques réalisées à Saint-Mihiel on associera les relevés de macroinvertébrés de Sassey-sur-Meuse.

FIGURE 2.6 – Carte situant les stations choisies

3.2 Groupement des données

Les données concernant les macroinvertébrés sont disponibles sur deux tableaux : l'un contenant les stations belges et l'autre les stations françaises. Avant de pouvoir travailler sur ces données il faut donc rassembler ces deux tableaux en un seul contenant toutes les données relatives aux macroinvertébrés sur chacune des stations étudiées.

Il faut prendre en compte que certaines espèces sont recensées sur certaines stations mais pas sur d'autres. Il faut donc les introduire dans le tableau où elles ne sont pas reprises en mettant leurs effectifs à zéro.

Dans les tableaux de données, les premières colonnes reprennent la classification scientifique de chaque espèce (embranchement, classe, ordre, famille et genre), les colonnes suivantes comprennent le nombre d'individus de chaque espèce pour une station et une année donnée. Par exemple :

embranchement	classe	ordre	famille	genre	Inor 2000	Inor 2001
crustacés	malacostracés	amphipodes	gammaridae		688	1976
insecte		ephemeropteres	caenidae	caenidae	288	1105

Or dans les tableaux de départ, une même colonne pouvait contenir deux types différents (par exemple la classe d'une espèce et l'ordre d'une autre). Les tableaux ont donc dû être réorganisés.

Une fois les stations belges et françaises rassemblées dans le même tableau, celui-ci contient 515 genres de macroinvertébrés différents. Travailler sur un si grand nombre de taxons serait fastidieux, et rendrait des résultats trop précis pour notre étude. En effet, nous étudions une évolution globale de la faune suite à l'apparition d'espèces de bivalves exogènes et non une évolution de chaque espèce. Les macroinvertébrés ont donc été regroupés en 22 taxons, qui sont des ordres dans la classification scientifique : les bivalvia, bryozoa, cnidaria, coleoptera, crustacea, diptera, ephemeroptera, gastropoda, hemiptera, hirudinea, hydracari, lepidoptera, megaloptera, nemathelminthes, nemertae, odonata, oligochaeta, planipennia, plecoptera, porifera, trichoptera et turbellaria. Afin de mieux appréhender l'augmentation des bivalves exogènes, la classe bivalvia est divisée en 2 classes : bivalvia « natifs »et bivalvia « invasifs ».

3.3 Manque d'informations

Les données reprennent un nombre représentant le nombre d'individus d'une certaine espèce chaque année dans chaque station, or le recensement n'a pas été fait chaque année, ce qui a créé des trous dans nos données. La gestion de ce manque de données sera explicitée dans le chapitre 4.

Chapitre 3

Tests de tendance

1 Non stationnarité des séries temporelles

« Une série temporelle, aussi appelée série chronologique, est une suite de valeurs numériques représentant l'évolution d'une quantité spécifique au cours du temps. »[29] Les séries temporelles sont souvent utilisées en biologie afin d'analyser le comportement de divers paramètres physiques, chimiques ou biologiques. Lors de l'étude de telles séries, une des questions qui revient couramment est de savoir si les données suivent un processus stationnaire. Cela signifie que l'on souhaite vérifier si la structure des données évolue ou non avec le temps. Si la série de données est stationnaire, sa moyenne n'évolue pas avec le temps. Les causes de non stationnarité peuvent être les suivantes [30] :

- une modification graduelle de la série au cours du temps, qui se manifeste par une baisse (ou hausse) des valeurs de la série,
- une rupture survenant à un temps donné, à partir de laquelle les caractéristiques de la série changent,
- un changement de la loi suivie par une variable à partir d'une date donnée.

Afin de tester la stationnarité d'une série, nous pouvons donc rechercher la présence d'une tendance. Une tendance correspond à une orientation prise par une série de données en fonction du temps. Il s'agit donc d'une évolution au cours du temps (que ce soit croissante ou décroissante).

Comme tout test statistique, le test de tendance teste une hypothèse nulle contre une hypothèse alternative. La statistique du test permet de décider si l'hypothèse nulle est acceptée ou rejetée.

Le test de tendance a pour hypothèse nulle H_0 l'absence de tendance dans la série et pour hypothèse alternative H_a la présence d'une tendance croissante ou décroissante.

La statistique du test dépend du test effectué, elle permet de choisir une des deux hypothèses. En statistique, l'hypothèse choisie est celle qui a le plus de chance d'être vraie, et non nécessairement celle qui *est* vraie. Par conséquent une hypothèse ne sera jamais « acceptée », on dira plutôt qu'on rejette ou retient l'hypothèse H à un niveau de confiance α. Le *niveau de confiance* α désigne « un seuil de probabilité donné, comparé à la valeur calculée de la "statistique" du test pour savoir si l'écart observé est compatible avec l'hypothèse nulle ou non. »[29]. Le résultat du test est dit *significatif* s'il est improbable qu'il soit obtenu par un simple hasard. Si une hypothèse H est retenue au niveau 0.05, par exemple, cela signifie que le résultat observé a moins de 5 % de chances d'être obtenu par hasard.

La p-valeur est la « probabilité d'obtenir la même valeur (ou une valeur encore plus extrême) de la statistique du test si l'hypothèse nulle était vraie. »[33]. C'est le niveau de signification le plus bas pour laquelle l'hypothèse nulle peut être rejetée. La règle de décision est la suivante :

$$\begin{cases} \text{p-valeur} > \alpha & \text{hypothèse nulle « acceptée »,} \\ \text{p-valeur} \leq \alpha & \text{hypothèse nulle rejetée.} \end{cases}$$

2 Les autocorrélations présentes dans les séries temporelles

Les séries de données présentent ce qu'on appelle des *autocorrélations* si une donnée observée à une date dépend des données antérieures.

Un *processus stochastique* est une fonction qui représente l'évolution d'une variable aléatoire [24]. La covariance entre deux variables x et y, notée $Cov(x, y)$ permet de quantifier l'écart par rapport à l'espérance respective de chacune de ces variables. La fonction d'autocovariance d'un processus stochastique permet de caractériser les dépendances linéaires existant au sein de ce processus [11]. Soit le processus $X = \{x_i, 1 \leq i \leq n\}$ à valeur dans $\mathbb{R}$. Si il admet une variance $V(x_i)$ pour tout i, la fonction d'autocovariance de X, notée R, est définie par

$$R(i, j) = Cov(x_i, x_j) = E\left[(x_i - E(x_i))(x_j - E(x_j))\right].$$

Si X est un processus stationnaire, alors l'espérance est constante au cours du temps, on a donc $E(x_i) = E(x_j)$. Dans ce cas $R(i, j) = R(|i-j|, 0)$ et les autocovariances peuvent se définir par la fonction qui associe à tout i $R(i, 0)$.

L'autocorrélation est obtenue en divisant l'autocovariance par la variance [10]. L'autocorrélation d'une série temporelle discrète ou d'un processus X est la corrélation du processus par rapport à une version décalée dans le temps de ce même processus. Si le processus est un processus stationnaire dont l'espérance vaut μ alors l'autocorrélation est définie par

$$R(k) = \frac{E(x_i - \mu)(x_{i+k} - \mu)}{\sigma^2},$$

où k est le décalage temporel et σ^2 la variance.

3 Présentation de différents tests de tendance

Différents tests existent afin de détecter la présence d'une tendance significative dans une série temporelle. Quelques uns d'entre eux sont présentés dans cette section. Ces tests de tendance sont des tests d'hypothèse, il existe deux types de tests : les tests paramétriques et non paramétriques. Les *tests paramétriques* ont pour hypothèse que les données, ainsi que leurs erreurs, suivent une loi donnée (généralement une loi normale). Cette condition doit être remplie pour pouvoir appliquer le test. Ils sont généralement plus puissants que les tests non-paramétriques. Les *tests non-paramétriques* ne nécessitent pas d'hypothèse sur la distribution des données.

3.1 La régression linéaire

Une régression linéaire a pour but d'ajuster les données par une droite. Il s'agit d'un test paramétrique qui nécessite que les erreurs suivent elles aussi une loi normale, que leur moyenne soit nulle et qu'elles ne soient pas corrélées entre elles.

La régression permet de vérifier s'il existe une tendance linéaire entre le temps (t) et les données (x). La fonction de régression est de la forme $x = \beta_0 + \beta_1 t$, où $\beta_0, \beta_1 \in \mathbb{R}$.

La pente de la droite de régression est calculée comme

$$\beta_1 = \frac{\sum(t_i - \bar{t})(x_i - \bar{x})}{\sum(t_i - \bar{t})^2},$$

et l'ordonnée à l'origine par :

$$\beta_0 = \bar{x} - \beta_1 \bar{t},$$

où $\bar{x}$ et $\bar{t}$ sont la moyenne des x_i et t_i [27], [40]. La statistique du test est, si le temps évolue de façon constante (on ajoute à chaque fois la même valeur au temps t_i précédent pour obtenir le temps suivant t_{i+1})

$$S = \beta_1/\sigma, \text{ où } \sigma = \sqrt{\frac{12\sum_{i=1}^{n}(x_i - \beta_0 - \beta_1 t_i)}{n(n-2)(n^2-1)}},$$

sinon

$$\sigma^2 = \frac{1}{n-2}\frac{\sum_{i=1}^{n}(x_i - (\beta_1 t_i + \beta_0)}{\sum_{i=1}^{n}(t_i - \bar{t})^2}.$$

Cette statistique suit une loi de Student à $n-2$ degrés de liberté sous l'hypothèse nulle. Les valeurs critiques de rejet de l'hypothèse nulle sont reprises dans les tables statistiques pour une loi de Student t pour différents niveaux de confiance.

3.2 Le test de Cox-Stuart

Il s'agit d'un test non paramétrique visant à rechercher une tendance dans une série de données. La série (de taille n) est divisée en deux parties en sa moitié. Des paires de données $(x_i, x_{i+n/2})$ sont créées et les termes de ces paires sont comparés. Si $x_{i+n/2} > x_i$ le signe « + »est associé à la paire, inversement si $x_{i+n/2} < x_i$ le signe « − »est associé à la paire.

Sous hypothèse nulle (absence de tendance) le nombre de paires positives suit une distribution binomiale dont la probabilité de succès ou d'échec est de 1/2. La probabilité d'obtenir n^- succès parmi $n^- + n^+$ épreuves est calculée, l'hypothèse nulle est rejetée si cette probabilité est inférieure au seuil de rejet α.

3.3 Le test de de Mann-Kendall modifié par Hamed et Rao

Cette méthode est basée sur les publications [43] et [49], le test de Mann-Kendall détermine si une tendance est identifiable dans les données, la modification apportée par Hamed et Rao sert à prendre en compte les autocorrélations présentes dans les données. En effet, dans des données représentant l'évolution d'une population ou d'un paramètre au cours du temps des autocorrélations peuvent apparaitre car une variable à une date donnée n'est pas indépendante de cette même variable à la date précédente.

Test de Mann-Kendall

La rédaction de cette partie est basée sur les documents [43] et [39] et les sites web [5], [7] et [46].

Le test de tendance de Mann-Kendall a pour but de tester les hypothèses suivantes :

$$\begin{cases} H_0 & : \text{ pas de tendance monotone} \\ H_1 & : \text{ tendance monotone présente} \end{cases}$$

Nous disposons de plusieurs séries temporelles de dimension n. Soit une de ces séries : $x_1, x_2, \ldots, x_n$ qui sont les mesures obtenues pour les temps $1, 2, \ldots, n$. Le test de Mann-Kendall calcule la statistique S :

$$S = \sum_{i<j} sgn(x_j - x_i),$$

où

$$sgn(x_j - x_i) \begin{cases} 1 & \text{si} \quad x_i < x_j, \\ 0 & \text{si} \quad x_i = x_j, \\ -1 & \text{si} \quad x_i > x_j. \end{cases}$$

Si S est positif, les observations tardives tendent à être plus grandes que les premières (la tendance est croissante), si S est négatif elles tendent à être plus petites que les premières observations (tendance décroissante).

Si $n \leq 10$:

Plusieurs tests d'hypothèse sont possibles :

1. $$\begin{cases} H_0 : \text{pas de tendance monotone} \\ H_1 : \text{tendance monotone croissante} \end{cases}$$

   H_1 est acceptée si $S > 0$ et si la probabilité correspondante dans la table 3.1 est inférieure au seuil de rejet α.

2. $$\begin{cases} H_0 : \text{pas de tendance monotone} \\ H_1 : \text{tendance monotone décroissante} \end{cases}$$

   H_1 est acceptée si $S < 0$ et si la probabilité correspondante dans la table 3.1 est inférieure au seuil de rejet α.

3. $$\begin{cases} H_0 : \text{pas de tendance monotone} \\ H_1 : \text{tendance monotone} \end{cases}$$

   H_1 est acceptée si le double de la probabilité correspondante dans la table 3.1 est inférieure au seuil de rejet α.

Si n ne peut être trouvé dans la table de probabilités (table 3.1), la valeur suivante sera utilisée. (Par exemple si on cherche $S = 12$ et qu'il n'y a pas de valeur dans la table, on prendra $S = 13$).

Si $n > 10$:

La variance de S est obtenue par

$$Var(S) = \frac{n(n-1)(2n+5)}{18}.$$

La statistique du test Z est calculée comme

$$Z = \begin{cases} \frac{S-1}{\sqrt{(Var(S))}} & \text{si} \quad S > 0, \\ 0 & \text{si} \quad S = 0, \\ \frac{S+1}{\sqrt{(Var(S))}} & \text{si} \quad S < 0. \end{cases} \tag{3.1}$$

Une valeur positive (négative) de Z indique une tendance croissante (décroissante) avec le temps. Plusieurs tests sont possibles pour un seuil d'erreur α :

1. $$\begin{cases} H_0 : \text{pas de tendance monotone} \\ H_1 : \text{tendance monotone croissante} \end{cases}$$

   H_1 est acceptée si $Z \geq Z_{1-\alpha}$,

2. $$\begin{cases} H_0 : \text{pas de tendance monotone} \\ H_1 : \text{tendance monotone décroissante} \end{cases}$$

   H_1 est acceptée si $Z \leq -Z_{1-\alpha}$,

3. $$\begin{cases} H_0 : \text{pas de tendance monotone} \\ H_1 : \text{tendance monotone} \end{cases}$$

   H_1 est acceptée si $|Z| \geq Z_{1-\alpha/2}$,

TABLE 3.1 – Table des probabilités que la statistique S soit plus grande ou égale que la valeur de S spécifiée [39]. Notation utilisée : $0^2 \equiv 00$.

	valeurs de n					valeurs de n		
S	4	5	8	9	S	6	7	10
0	0.625	0.592	0.548	0.540	1	0.500	0.500	0.500
2	0.375	0.408	0.452	0.460	3	0.360	0.386	0.431
4	0.167	0.242	0.360	0.381	5	0.235	0.281	0.634
6	0.042	0.117	0.274	0.306	7	0.136	0.191	0.300
8		0.042	0.199	0.238	9	0.068	0.119	0.242
10		0.0^283	0.138	0.179	11	0.028	0.068	0.190
12			0.089	0.130	13	0.0^283	0.035	0.146
14			0.054	0.090	15	0.0^214	0.015	0.108
16			0.031	0.060	17		0.0^254	0.078
18			0.016	0.038	19		0.0^214	0.054
20			0.0^271	0.022	21		0.0^320	0.036
22			0.0^228	0.012	23			0.023
24			0.0^387	0.0^263	25			0.014
26			0.0^319	0.0^229	27			0.0^283
28			0.0^425	0.0^212	29			0.0^246
30				0.0^343	31			0.0^223
32				0.0^312	33			0.0^211
34				0.0^425	35			0.0^347
36				0.0^528	37			0.0^318
					39			0.0^458
					41			0.0^415
					43			0.0^528
					45			0.0^628

où Z_α est le $100\alpha^{\text{e}}$ percentile de la distribution d'une loi normale. Pour $\alpha = 0.05$, $Z_{1-\alpha/2} = 1.96$.

Modification de Hamed et Rao

Lors de l'étude de séries temporelles, une autocorrélation existe entre les variables. Cette dépendance entre les variables peut conduire, à tort, à la conclusion d'une tendance significative. Pour prendre en compte ces autocorrélations temporelles il faut corriger la variance des variables.

Une régression linéaire est effectuée sur les données (x) ayant subit la transformation logarithmique $\ln(x+1)$. Afin de se débarrasser des autocorrélations, une estimation non paramétrique des variables, évaluée par cette régression, est soustraite de ces données. Le test de tendance de Mann-Kendall est ensuite appliqué sur ces nouvelles données.

Pour prendre en compte ces autocorrélations temporelles, la variance est corrigée comme tel :

$$Var^*(S) = Var(S) * Cor, \tag{3.2}$$

où

$$Cor = 1 + \frac{2}{n(n-1)(n-2)} \sum_{i=1}^{n} (n-1)(n-i-1)(n-i-2)\rho_S(i),$$

et $\rho_S(i)$ est le coefficient d'autocorrélation significativement différent de zéro au lag i. Les coefficients $\rho_S(i)$ sont donnés pour chaque pas avant qu'ils ne soient trop proche de 0.

3.4 Block bootstrapping basé sur le test de Mann Kendall

Cette section est basée sur le document [37], et présente une autre manière d'utiliser le test de Mann Kendall tout en prenant compte des autocorrélations.

Soit un ensemble de données $X = (x_1, \ldots, x_n)$, dont la statistique de Mann Kendall est S. La significativité de S peut être évaluée en se basant sur la distribution de la statistique de Mann Kendall bootstrappée, $BECD \sim \hat{S}^*$, qui est dérivée de l'échantillon de données « bootstrappé ».

Un échantillon « bootstrappé », noté $X^* = (x_1^*, \ldots, x_n^*)$ est obtenu en échantillonnant aléatoirement n fois avec remplacement et avec probabilité égale $1/n$ l'échantillon observé $x_1, \ldots, x_n$. En bootstrappant l'échantillon X un nombre M de fois, on peut obtenir M échantillons bootstrappés indépendants $X^{*1}, \ldots, X^{*M}$ chacun de taille n. La statistique S de Mann Kendall est calculée pour chaque échantillon bootstrappé, on obtient ainsi M statistiques $\hat{S}^* = (S^{*1}, \ldots, S^{*M})$. En les ordonnant de façon croissante, la distribution empirique de la statistique ($BECD \sim \hat{S}^*$) sera obtenue. La méthode utilisée consiste à rééchantillonner les données par blocs de longueur fixée l, cette longueur de bloc doit être ajustée en fonction de la variable à lequel le block bootstrapping est appliqué.

P-valeur

La p-valeur (p_S) de S de l'échantillon observé est estimée en utilisant la courbe $BECD \sim \hat{S}^*$ par

$$p_S = P(\hat{S}^* \leq S) = \frac{m_S}{M}$$

où m_S est le rang correspondant à la plus grande valeur $\hat{S}^*$ telle que $\hat{S}^* \leq S$.

Pour un échantillon ne comportant aucune tendance, la p-valeur devrait être proche de 0.5. Une statistique S positive ou négative correspond, respectivement, à une tendance croissante ou décroissante. Au seuil de rejet $\alpha = 0.05$ pour un test unilatéral :

$$\begin{cases} p_S \leq 0.05 : \text{tendance négative significative} \\ p_S \geq 0.95 : \text{tendance positive significative} \end{cases}$$

Intervalle de confiance

Soient un échantillon aléatoire simple $X_1, \ldots, X_n$ d'une loi de densité $f(x, \theta)$; $T_1(X_1, \ldots, X_n)$ et $T_2(X_1, \ldots, X_n)$ deux statistiques telles que $P(T_1 \leq \theta \leq T_2) = \gamma$ où θ et γ sont indépendants. Alors l'intervalle aléatoire $[T_1, T_2]$ est appelé intervalle de confiance à $\gamma.100\%$ [34] . Environ $\gamma.100\%$ de ces intervalles contiennent la vraie valeur de θ.

La méthode du percentile est utilisée afin de construire les intervalles de confiance. Pour un test bilatéral de seuil $\alpha = 1 - \gamma$, la méthode du percentile est l'intervalle compris entre les percentiles $100.\alpha/2$ et $100.(1 - \alpha/2)$ de la distribution bootstrap de S^*. Le $100.\alpha/2$ percentile de la distribution bootstrap de S^* est estimé en ordonnant S^* par ordre croissant, ensuite en interpolant entre les éléments ordonnés $(\alpha M/2)$ et $(\alpha M/2+1)$ de S^*. Si le nombre d'échantillons bootstrappés, M, est assez grand, on obtient un intervalle de confiance précis par cette méthode. Pour un intervalle de confiance 90-95% Davison et Hinkley (1997) suggèrent que M devrait être entre 1000 et 2000. Nous utiliserons dans nos tests $M = 1500$.

4 Test de normalité des données

Différents tests existent afin de vérifier la normalité des données. Ceux-ci sont régulièrement utilisés par les statisticiens afin de vérifier si l'application d'un test paramétrique est possible.

Le test le plus connu est celui de **Kolmogorov-Smirnov**. Il est utilisé pour déterminer si une échantillon suit une loi dont on connait la fonction de répartition ou si deux échantillons suivent la même loi.

Soit $(x_1, \ldots, x_n)$ une échantillon de variables aléatoires indépendantes, la fonction de répartition empirique de cet échantillon est définie par [31] :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i \leq x},$$

où $\delta_{x_i \leq x} = \begin{cases} 1 & \text{si} \quad xi \leq x, \\ 0 & \text{si} \quad \text{sinon.} \end{cases}$

L'hypothèse nulle est que les deux échantillons proviennent de la même distribution. La statistique du test est basée sur la différence en valeur absolue entre la fonction de répartition empirique de l'échantillon et une fonction de répartition connue ou la répartition empirique du deuxième échantillon dans le cadre d'une comparaison de deux échantillons. Le choix du rejet ou non de l'hypothèse nulle se fait à partir de la table de Smirnov [50].

Le **test d'Anderson-Darling** est une modification du test de Kolmogorov-Smirnov adaptée à plusieurs lois. Cette méthode est applicable à une loi normale dont on ne connait pas les paramètres.

Le **test de Shapito-Wilk** [32] est adapté pour de petits échantillons. Il teste l'hypothèse nulle selon laquelle l'échantillon est issu d'une population normalement distribuée. Soit un échantillon $x_1, \ldots, x_n$, la statistique du test est :

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

où $x_{(i)}$ désigne la statistique d'ordre (les x_i ordonnés par ordre croissant) et $\bar{x}$ la moyenne de l'échantillon. Les constantes a_i sont données par

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}},$$

où $m = (m_1, \ldots, m_n)^T$ et m_i est l'espérance des statistiques d'ordre d'un échantillon de variables indépendantes et identiquement distribuées suivant une loi normale. La matrice V est la matrice de variance-covariance de ces statistiques d'ordre. L'hypothèse nulle est rejetée si W est « trop petit », c'est à dire si la p-valeur du test est inférieure au seuil de rejet α.

Chapitre 4

Application de tests de tendance aux données

Le but de ce chapitre est de détecter les tendances (croissantes ou décroissantes) existantes dans les données reprenant les paramètres physicochimiques des stations étudiées. Une chute significative du taux de chlorophylle a indiquant une diminution de la population de phytoplancton, nous étudions le comportement des différents paramètres physicochimiques au cours du temps.

Suite à la lecture de plusieurs publications ([43] et [49]) il ressort que le test généralement appliqué en biologie est celui de Mann Kendall. Dans un premier temps, le test de Mann-Kendall modifié par Hamed et Rao est appliqué, dans un second temps nous utilisons le block bootstrapping basé sur le test de Mann Kendall. Une comparaison est ensuite réalisée afin de choisir la méthode la plus adaptée à notre étude.

1 Modification apportées aux données

Avant d'appliquer les tests aux données, il est nécessaire de leur apporter quelques modifications. D'une part les données sont des séries temporelles incomplètes, or, les tests utilisés ne tolèrent pas de manque de données dans la série. D'autre part, des erreurs de mesure sont présentes dans les séries temporelles, qui comportent donc ce que l'on appelle des outliers [1], il faut alors les traiter afin de ne pas induire d'erreurs dans les calculs.

1.1 Complétion des séries temporelles

Les séries temporelles contenant les données ne sont pas complètes. En effet, certaines années, les paramètres physicochimiques de l'eau n'ont pas été enregistrés. Pour palier ce manque, une droite est tracée entre les données possédées (comme présenté en figure 4.1). Entre deux données $(an_1, param_1)$ et $(an_2, param_2)$, la pente de la droite d'équation $y = mx + p$ est

$$m = \frac{(param_2 - param_1)}{(an_2 - an_1)},$$

et

$$p = param_2 - m.an_2.$$

Cette méthode permet d'obtenir des séries temporelles complètes. Les données manquantes au début et à la fin des séries ne seront pas considérées.

1. En statistique, un outlier est une observation qui est distante des autres [20].

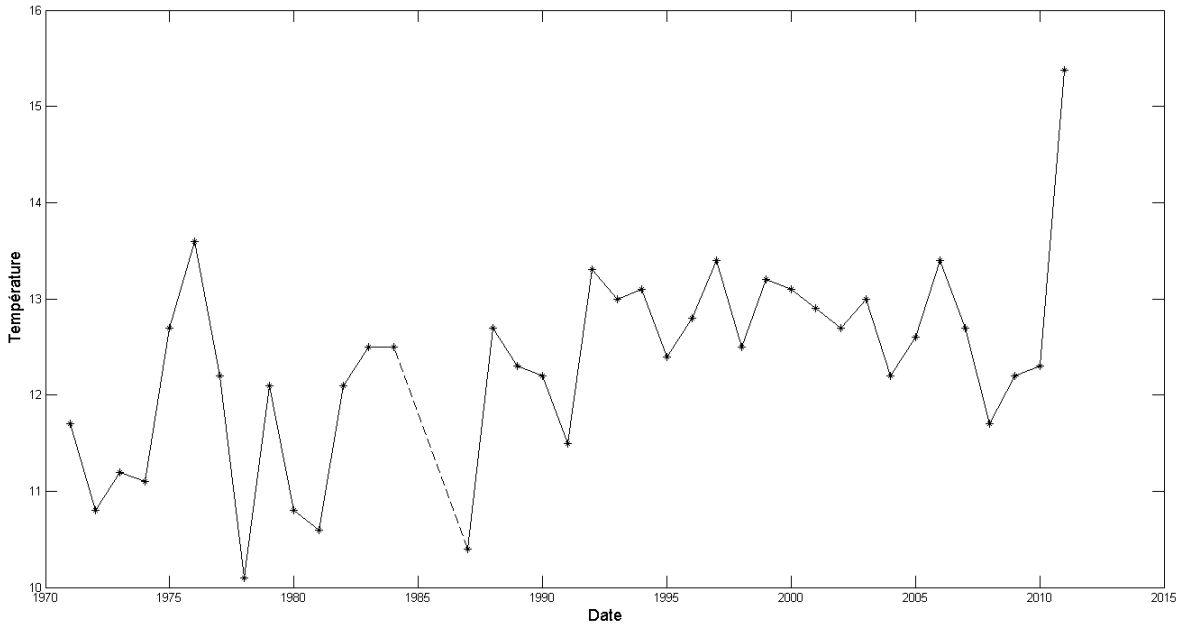


FIGURE 4.1 – Exemple de complétion des données, la courbe en trait plein représente les données disponibles et la courbe en pointillé les données créées pour compléter la série temporelle

1.2 Prise en compte des outliers

Retrait des outliers

L'algorithme de retrait des outliers consiste à trouver un potentiel outlier, dont la valeur est la plus éloignée de la moyenne de l'échantillon. Un test est appliqué à cette valeur afin de vérifier s'il s'agit d'un outlier significatif pour les données. Si un outlier $x_i, (1 \leq i \leq n)$ est détecté (avec une p-valeur inférieure à 0.05) il est remplacé par

$$x(i) = \begin{cases} \frac{x_{i-1}+x_{i+1}}{2} & \text{si} \quad 1 < i < n, \\ x(n-1) & \text{si} \quad i = n, \\ x(2) & \text{si} \quad i = 1. \end{cases}$$

Tant que des outliers significatifs sont détectés, un autre outlier est recherché.

La figure 4.2 présente un exemple de données auxquelles les outliers ont été détectés et remplacés.

Différentes méthodes permettant de vérifier la significativité des outliers existent, entre autres : les méthodes de Dixon, Cochran, Grubbs et Chi-carré.

Le test de Dixon [44],[17] consiste à comparer la distance entre les points les plus éloignés et les points immédiatement voisins à l'étendue totale des résidus pour des données suivant une loi normale. Il est conseillé de ne pas utiliser plusieurs fois ce test dans le même jeu de données.

Le test de Cochran [16] compare si une variance est significativement plus grande que les autres. Ce test suppose lui aussi la normalité des données.

Le test de Grubbs [19] détecte les valeurs aberrantes en terme de dispersion de moyennes à partir de données dont la distribution est normale. Ce test est beaucoup plus puissant que le test de Dixon pour de petits échantillons.

Le test du chi-carré [13] effectue un test basé sur la distribution chi-carré du carré des différences entre les données et la moyenne de l'échantillon.

Ces tests ont pour hypothèse nulle que l'observation considérée n'est pas un outlier. La valeur considérée sera donc un outlier si la p-valeur du test est inférieure au seuil de rejet 0.05.

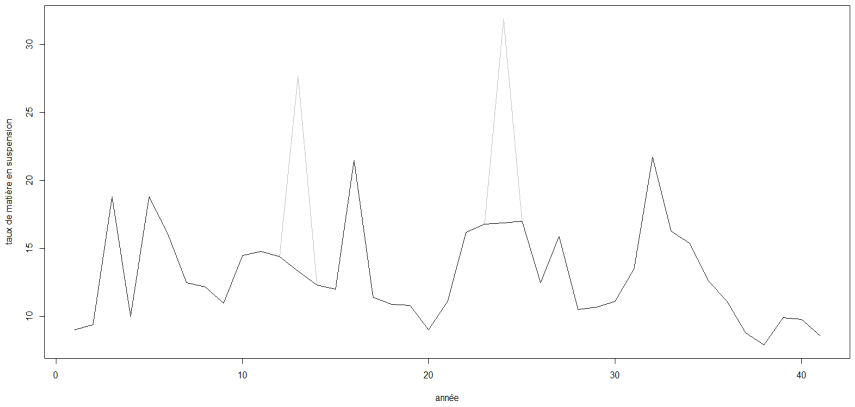


FIGURE 4.2 – Taux de matière en suspension par rapport au temps à Saint Mihiel, la courbe grise représente les données avec outliers et la courbe noire celles sans outliers

Vérification de la normalité des données

La plupart des méthodes de détection des outliers nécessitent la normalité des données. Pour vérifier cette hypothèse un test de Shapiro est effectué ainsi qu'un diagramme quantile-quantile.

Le test de Shapiro [32] teste l'hypothèse nulle selon laquelle l'échantillon est issu d'une population normalement distribuée. Soit un échantillon $x_1, \ldots, x_n$, la statistique du test est :

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

où $x_{(i)}$ désigne la statistique d'ordre (les x_i ordonnés par ordre croissant) et $\bar{x}$ la moyenne de l'échantillon. Les constantes a_i sont données par

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}},$$

où $m = (m_1, \ldots, m_n)^T$ et m_i est l'espérance des statistiques d'ordre d'un échantillon de variables indépendantes et identiquement distribuées suivant une loi normale. La matrice V est la matrice de variance-covariance de ces statistiques d'ordre. L'hypothèse nulle est rejetée si W est « trop petit », c'est à dire si la p-valeur du test est inférieure au seuil de rejet α (fixé à 0.05).

La normalité des données peut aussi être vérifiée par un graphique quantile-quantile (QQ-plot), les données suivent une loi normale si les points forment une droite sur le graphe.

Application aux données

Le test de Shapiro est appliqué à la variable *Ptot* sur la station de Han-sur-Meuse, la p-valeur rendue est $1.55e-08$, l'hypothèse nulle selon laquelle les données suivent une loi normale est donc rejetée. Le diagramme quantile-quantile (figure 4.3) confirme le test de Shapiro ; en effet, si les données suivaient une loi normale elles se situeraient le long de la droite centrale, entre les courbes supérieure et inférieure représentant l'intervalle de confiance.

En observant l'histogramme de cette même variable, présenté en figure 4.4, les variables semblent suivre une loi exponentielle, ce qui se confirme par le QQplot en figure 4.5. Le test de Grubbs ne peut donc pas être appliqué à toutes les variables, pour cette raison, le test de chi-carré, plus général, sera utilisé.

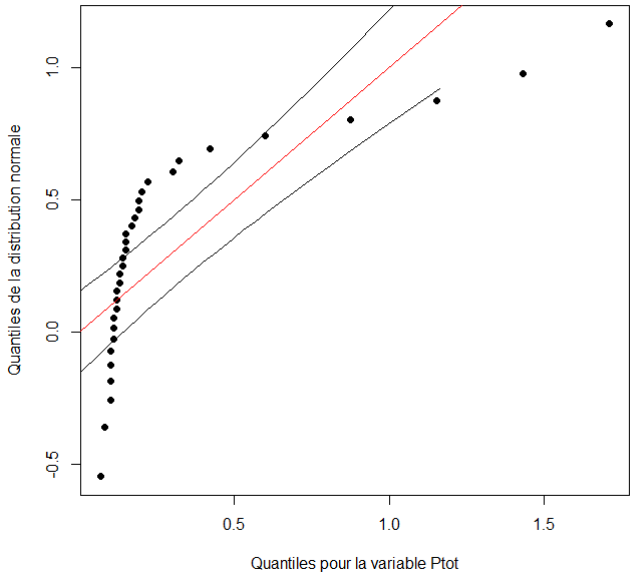


FIGURE 4.3 – Diagramme quantile quantile pour la variable *Ptot*, la droite au centre représente le chemin suivi par des données suivant une loi normale, les courbes supérieure et inférieure représentent l'intervalle de confiance

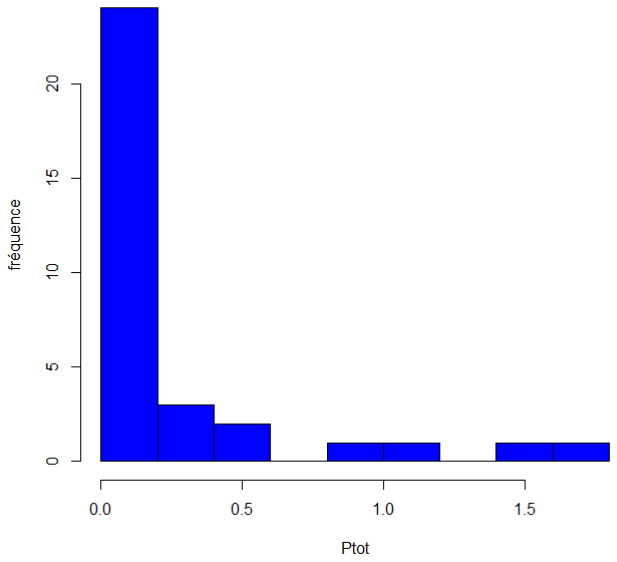


FIGURE 4.4 – Histogramme de la variable *Ptot* pour la station Han-sur-Meuse

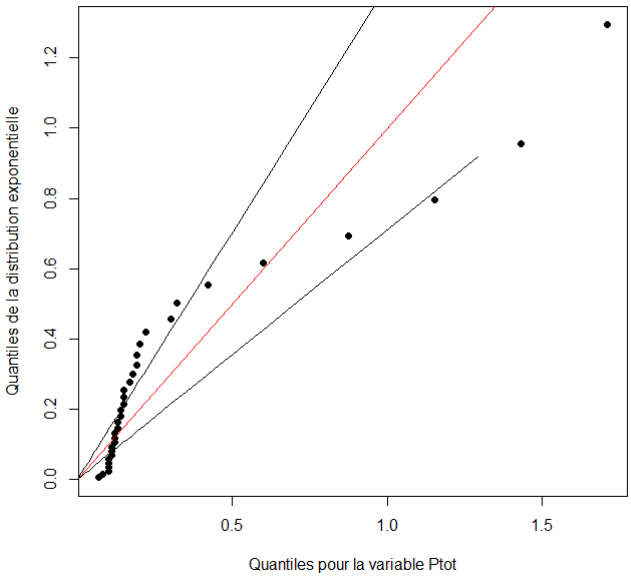


FIGURE 4.5 – Diagramme quantile quantile pour la variable *Ptot*, la droite centrale représente le chemin suivi par des données suivant une loi exponentielle, les courbes supérieure et inférieure représentent l'intervalle de confiance

Après application sur les différentes variables, il s'avère que l'algorithme de retrait des outliers boucle pour les variables NO_3 sur la station de Saint Mihiel et *Ptot* sur les stations de Han-sur-Meuse et Inor. La normalité de ces données est vérifiée par le test de Shapiro et un QQplot. Le test de Shapiro indique une p-valeur de 0.04853 pour la variable NO_3 à Saint-Mihiel. Le QQplot en figure 4.6 montre néanmoins que les données suivent une loi normale, le test de Grubbs est appliqué car, de tous les tests présentés nécessitant la normalité des données, c'est celui qui est le plus adapté à nos données (petit jeu de données auquel on veut appliquer le test plusieurs fois). La même analyse est effectuée pour la variable *Ptot* et la même conclusion en est faite : la variable peut être considérée normale et le test de Grubbs est appliqué.

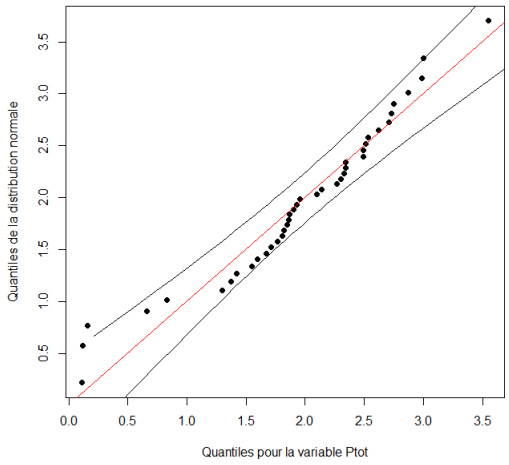


FIGURE 4.6 – Diagramme quantile quantile pour la variable NO_3, la droite au centre représente le chemin suivi par des données suivant une loi normale, les courbes supérieure et inférieure représentent l'intervalle de confiance

2 Application des tests de tendance

2.1 Test de Mann Kendall modifié par Hamed et Rao

Les résultats de ce test sont présentés dans la table 4.1. Peu de tendances significatives sont relevées, 5 variables sur les 9 et sur certaines stations présentent une tendance significative. Après discussion avec le commanditaire, le test ne révélant pas la réalité observée par les biologistes, nous décidons d'appliquer un test différent aux données.

Code R

Le code utilisé est en annexe 1. Le package `Kendall` est chargé, ce package contient les fonctions relatives aux tests de tendance de Kendall. Une variable physicochimique est ensuite chargée, et placée dans un vecteur différent pour chaque station. Le même travail est effectué pour chaque station : une nouvelle variable, ne contenant pas de valeurs indisponibles (`na` : not available) est créée via la fonction `na.omit`. Le test de Mann Kendall est appliqué à la variable contenant les `na` par la fonction `Kendall`, dont on extrait le score (variable S décrite en section 3.3) et la variance de ce score. Les autocorrélations et autocorrélations partielles sont tracées par les fonctions `acf` et `pacf`. La variance de S est corrigée par la formule (3.2) et la variable Z est ensuite calculée par l'équation (3.3).

TABLE 4.1 – Tendances des différents paramètres physicochimiques sur les stations étudiées par la méthode de Mann-Kendall modifée par Hamed et Rao

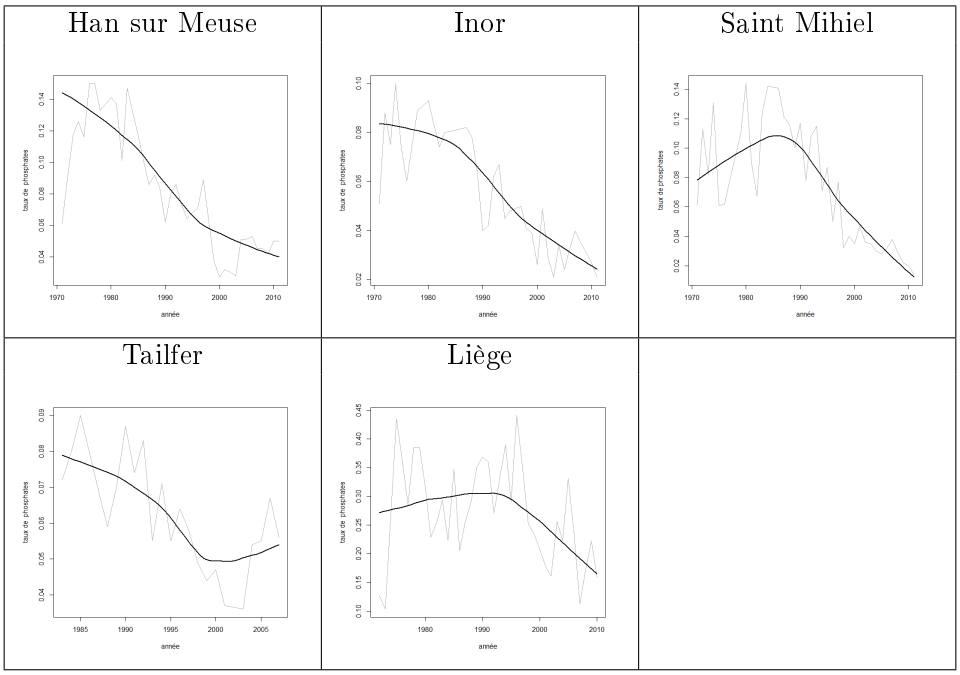
paramètre	station	S	Z	tendance
mes	Han-sur-Meuse	-119	-1.413093	_
	Inor	-279	-1.698448	_
	Saint-Mihiel	-61	-1.278586	_
	Tailfer	341	1.785152	_
	Liège	-59	-0.0811	_
Chla	Han-sur-Meuse	-215	-2.175581	↓
	Inor	-206	-2.185859	↓
	Saint-Mihiel	-134	-1.880724	_
	Tailfer	-150	-0.9905277	_
	Liège	-182	-1.10777	_
Amonium	Han-sur-Meuse	-119	-1.413093	_
	Inor	-116	-2.720176	↓
	Saint-Mihiel	-102	-2.01256	↓
	Tailfer	341	1.785152	_
	Liège	-59	-0.0811	_
Nitrate	Han-sur-Meuse	375	2.173844	↑
	Inor	529	2.240856	↑
	Saint-Mihiel	209	1.273007	_
	Tailfer	387	1.829043	_
	Liège	-338	-1.860295	_
phosphate	Han-sur-Meuse	-462	-1.868352	_
	Inor	-454	-1.865655	_
	Saint-Mihiel	-384	-1.757058	_
	Tailfer	93	0.6226632	_
	Liège	-126	-1.236984	_
Phosphore	Han-sur-Meuse	-396	-2.421043	↓
	Inor	-388	-2.367137	↓
	Saint-Mihiel	-106	-0.9085174	_
	Tailfer	33	0.349999	_
	Liège	93	0.9837002	_
Oxygène	Han-sur-Meuse	319	1.911555	_
	Inor	333	1.829292	_
	Saint-Mihiel	408	2.180357	↑
	Tailfer	349	2.080531	↑
	Liège	377	2.03064	↑
Débit	Han-sur-Meuse	5	0.0475034	_
	Inor	36	0.5011143	_
	Saint-Mihiel	-12	-0.1466318	_
	Tailfer	319	1.887756	_
	Liège	36	0.3577047	_
Température	Han-sur-Meuse	254	1.94733	_
	Inor	237	1.499007	_
	Saint-Mihiel	125	1.327209	_
	Tailfer	250	1.927212	_
	Liège	391	1.89226	_

2.2 Block bootstrapping appliqué au test de Mann Kendall

Suite aux résultats peu concluants de la méthode de Mann Kendall modifiée par Hamed et Rao, une nouvelle méthode est appliquée : le block bootstrapping appliqué au test de Mann Kendall.

Le détail des calculs menant à la détection de tendances croissantes ou décroissantes significatives est présenté en détail pour la variable représentant le taux de phosphate de l'eau (PO_4^3). Pour chaque station, l'évolution de la variable au cours du temps est tracées en gris (voir figure 4.7), la courbe lowess est tracée en gras. La courbe lowess est une courbe qui ajuste un nuage de points. Pour chaque point du nuage, elle calcule un polynôme de degré 2 qui ajuste ce point et les points environnants par une méthode des moindres carrés. Une pondération est effectuée, plus le point est éloigné du point de référence plus celle-ci est faible [21]. On peut détecter sur ces graphes la présence d'une tendance croissante ou décroissante, mais la significativité de celle-ci n'est pas connue.

FIGURE 4.7 – Courbe lowess pour les différents taux de phosphate sur les différentes stations



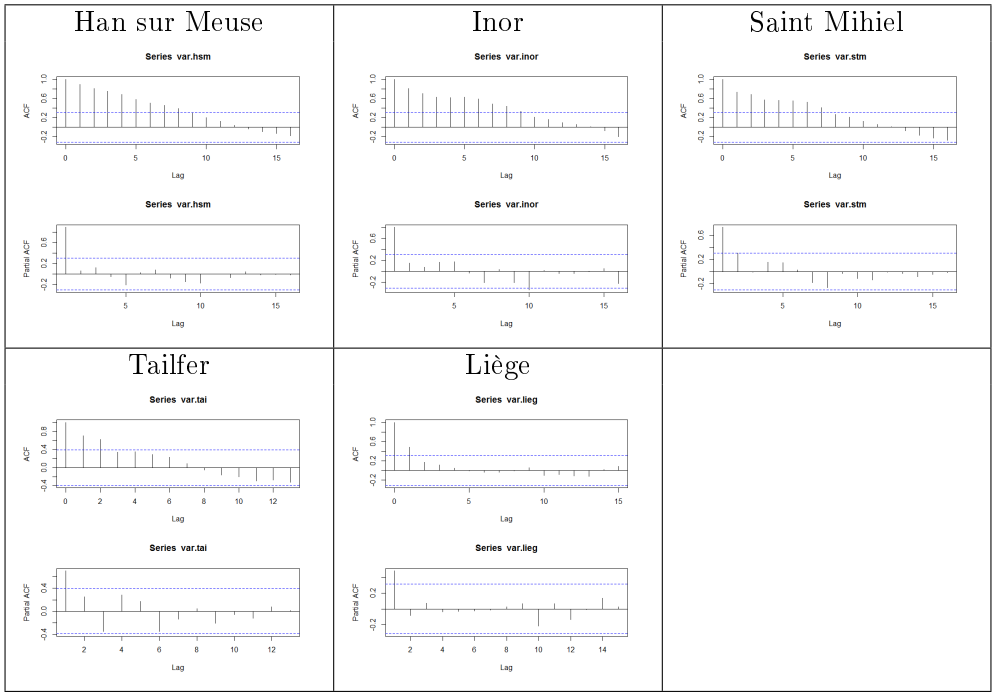
Avant d'appliquer le test de Mann Kendall et de conclure à des tendances significatives ou non, il faut vérifier si l'autocorrélation entre les données n'est pas significative. Pour cela, les autocorrélations et autocorrélations partielles des séries temporelles sont calculées (ACF et ACF partielle). Les graphes rendus sont présentés en figure 4.8. Les autocorrélations ne sont pas significatives si la plupart des pics verticaux de l'ACF et de l'ACF partielle sont compris entre les deux lignes horizontales en pointillés.

Pour la variable PO_4^3, les autocorrélations des stations de Tailfer et Liège ne sont pas significatives. Celles de Han-sur-Meuse, Inor et Saint-Mihiel le sont, il faudra donc appliquer le bootstrapping. Les résultat du test de Mann Kendall sont repris dans la table 4.2.

Le test effectué via le block bootstrapping étant basé sur des valeurs aléatoires, les p-valeurs rendues sont toutes différentes. Pour cette raison, la p-valeur et le niveau de significativité ne sont pas présentés dans la table 4.2. Afin d'obtenir des p-valeur qui ont toutes (ou du moins dont la majorité) la même signification pour le test, il faut réduire la longueur des blocs l. La longueur l (fixée à 5 par défaut) est strictement supérieure à 1 car pour $l = 1$ les autocorrélations ne sont pas prises en compte.

Le même travail est effectué pour les autres paramètres physicochimiques dont les graphes

FIGURE 4.8 – ACF complètes et partielles pour le phosphate sur les différentes stations



sont en annexe B. On observe la présence de beaucoup plus de tendances significatives par rapport au test modifié par Hamed et Rao, de plus ces tendances reflètent mieux ce qu'observent les biologistes.

Le niveau de significativité est décrit comme :

$$\begin{cases} * & \text{si} \quad \text{p-valeur} < 0.05 & \text{significatif} \\ ** & \text{si} \quad \text{p-valeur} < 0.01 & \text{hautement significatif} \\ *** & \text{si} \quad \text{p-valeur} < 0.001 & \text{très hautement significatif} \end{cases}$$

Nous observons une tendance négative significative de la chlorophylle a sur la plupart des stations. La chlorophylle a étant un bon indicateur de la biomasse de phytoplancton, elle est aussi liée au taux de matière en suspension, le phytoplancton faisant lui même partie des matières en suspension. Il n'est donc pas étonnant de remarquer qu'aux stations présentant une chute de la chlorophylle a, le taux des matières en suspension chute lui aussi. Sur la station de Han-sur-Meuse où la tendance descendante ($S = -74$) de chlorophylle a n'est pas significative, on remarque néanmoins une chute significative de matière en suspension. Notons que si la tendance n'est pas significative pour la chlorophylle a, la p-valeur est proche du seuil de 0.05 fixé. Sur la station de Saint-Mihiel où la chlorophylle a suit une tendance croissante significative, on remarque que la tendance décroissante du taux de matière en suspension est clairement rejetée.

La hausse du taux de nitrates ne peut être directement liée à la chute de phytoplancton. En effet l'impact humain est trop fort par rapport à cette variable. Par exemple les stations d'épuration rejettent de l'azote sous forme dégradée, parfois sous la forme NO^{3-}, les divers engrais utilisés en agriculture qui se retrouvent dans l'eau suite à l'érosion du sol apportent eux aussi des nitrates.

La variable mesurant le débit de la Meuse ne montre pas de tendance croissante, ce qui est normal car même si le débit peut varier en fonction des périodes de l'année, en moyenne annuelle celui-ci reste relativement semblable d'année en année.

TABLE 4.2 – Résultats des test de tendance avec block bootstrapping pour les paramètres physicochimiques sur les différentes stations

paramètre	station	S	p-valeur	tendance	
mes	Han-sur-Meuse	-242	0.0067706	**	↓
	Inor	-369	3.536e-05	***	↓
	Saint-Mihiel	-43	0.63699		_
	Tailfer	-92	0.010117	*	↓
	Liège	-210	0.0019026	**	↓
Chla	Han-sur-Meuse	-74	0.082541		_
	Inor	-85	0.045357	*	↓
	Saint-Mihiel	125	0.0018052	**	↑
	Tailfer	-287	0.00030411	***	↓
	Liège	-269	0.00071975	**	↓
Amonium	Han-sur-Meuse	-309	0.00048136	***	↓
	Inor	-196	0.027525	*	↓
	Saint-Mihiel	-155	0.081893		_
	Tailfer	-65	0.35855		_
	Liège	-185	0.025975	*	↓
Nitrate	Han-sur-Meuse	378	2.265e-05	***	↑
	Inor	558			_
	Saint-Mihiel	172	0.046306	*	↑
	Tailfer	144	0.00076306	***	↑
	Liège	-390			_
Phosphate	Han-sur-Meuse	-507			_
	Inor	-520			_
	Saint-Mihiel	-442			_
	Tailfer	-159	0.00082755	***	↓
	Liège	-181			↓
Phosphore	Han-sur-Meuse	-395			↓
	Inor	-387			↓
	Saint-Mihiel	-254			↓
	Tailfer	-33	0.36109		_
	Liège	-124	0.00051566	***	↓
Oxygène	Han-sur-Meuse	-76	0.062839		_
	Inor	-67	0.10151		_
	Saint-Mihiel	81	0.047149	*	↑
	Tailfer	234	0.0023005	**	↑
	Liège	-4	0.94044		_
Débit	Han-sur-Meuse	71	0.35301		_
	Inor	110	0.15398		_
	Saint-Mihiel	72	0.35301		_
	Tailfer	-58	0.18287		_
	Liège	-41	0.615		_
Température	Han-sur-Meuse	251	0.0048898	**	↑
	Inor	233	0.009027	**	↑
	Saint-Mihiel	127	0.14178		_
	Tailfer	220	0.0079373	**	↑
	Liège	343	3.3975e-05	***	↑

Code R

Le code utilisé est en annexe 2. Les packages `outlier`, `Kendall` et `boot` sont chargés, le premier contient les fonctions relatives à la détection des outliers, le second les fonctions permettant les tests de tendance de Kendall et le dernier les fonctions utilisées pour le block bootstrapping.

Une variable est chargée et mise dans un vecteur différent pour chaque station. Ces vecteurs sont transformés en séries temporelles par la fonction `ts`. Afin de procéder au remplacement des outliers, seules les valeurs disponibles sont prises en compte, la fonction `na.omit` est alors utilisée. L'algorithme de remplacement des outliers est mis en place : la p-valeur est initialisée à 0, tant que cette p-valeur sera inférieure au seuil de rejet (0.05), le programme identifiera des outliers via la fonction `outlier` et déterminera s'il s'agit d'outliers significatifs par la fonction `chisq.out.test` (`grubbs.test` pour les variables normales dont l'algorithme boucle avec le test de chi-carré). La p-valeur sera remplacée par la p-valeur rendue par le test chi-carré (ou le test de Grubbs), la boucle s'arrête quand plus aucun outlier significatif n'est détecté.

Suite au retrait des outliers, les fonctions `acf` et `pacf` permettent de tracer les graphiques relatifs aux autocorrélations. La courbe lowess est tracée par la fonction `scatter.smooth`.

Le test de Mann Kendall est appliqué aux données par la fonction `MannKendall`. Pour les données présentant une autocorrélation significative le block bootstrapping est appliqué. Une fonction calculant le tau, τ, de Kendall est créée. Le block bootstrapping est appliqué à cette fonction par `tsboot`. L'intervalle de confiance relatif à ce test est calculé par la fonction `boot.ci`, la p-valeur associée à cet intervalle est ensuite calculée par la fonction `t.test`. C'est à partir de cette p-valeur que la tendance sera dite significative ou non.

3 Comparaison des méthodes

Deux méthodes basées sur le test de Mann Kendall ont été étudiées et appliquées : la méthode modifiée par Hamed et Rao et le block bootstrapping. Elles prennent toutes deux en compte les autocorrélations présentes dans les séries temporelles. La première méthode ne prend pas en compte les outliers, les premiers résultats ne prenant pas compte des outliers ont été montrés au commanditaire. Suite à cela la méthode a été abandonnée car les résultats ne reflétaient pas les observations faites par le commanditaire. La deuxième méthode fait apparaitre beaucoup plus de tendances significatives avant la prise en compte des outliers et représente la réalité observée par le commanditaire. Le retrait des outliers a accentué ces tendances.

La différence majeure entre ces deux méthodes est que la première modifie les données avant de les utiliser afin de prendre en compte les autocorrélations, sans savoir si elles sont présentes ou non ; alors que la deuxième méthode vérifie la présence d'autocorrélations avant d'appliquer directement ou non le test de Mann Kendall. Si des autocorrélations sont présentes le block bootstrapping crée des jeux de données supplémentaires et augmente ainsi le nombre de données.

Chapitre 5

Méthodes de classification

La classification est une méthode statistique permettant de grouper les données selon différents critères. Elle permet d'une part de déterminer le nombre de classes présentes dans les données, mais aussi d'identifier quelle observation appartient à quelle classe.

1 Méthodes de choix du nombre de classes

Différentes méthodes permettent de déterminer le nombre de classes présentes dans les données, quelques-unes d'entre elles sont présentées dans cette section. Ces méthodes sont la méthode du dendrogramme et les méthodes basées sur les indices du déterminant CCC, du pseudo F et du t^2 et la méthode du coude. Les différentes méthodes sont présentées, accompagnées de la marche à suivre afin de déterminer le nombre de classes pour chacune d'entre elles. Les méthodes présentées dans cette section sont en majeures parties basées sur [41].

Soit un ensemble de données $(x_1, \ldots, x_n)$, une partition en k classes est recherchée. Il faut déterminer la valeur de k de façon à obtenir la partition optimale. Une partition d'un ensemble de données est une division de ces données en k classes, dont aucune n'est vide, dont l'intersection deux à deux est vide et dont l'union forme l'ensemble de données de départ.

Le critère de classification cubique (CCC)

L'indice du critère de classification cubique est défini comme

$$CCC = \ln\left[\frac{\frac{1-E(R^2)}{1-R^2}}{\frac{(np/2)^{1/2}}{(0.001+E(R^2))^{1.2}}}\right],$$

où n est le nombre d'observations, p est l'estimation de la « dimensionnalité »(déterminée par une analyse en composantes principales) et R^2 est le pourcentage de la variance expliquée par les classes. L'espérance $E(R^2)$ est calculée en supposant que les données sont générées suivant une distribution uniforme multidimensionnelle.

La valeur de l'indice correspondant au pic maximal indique le nombre de classes à retenir. La classification est bonne si $CCC > 2$ ou 3. Si $0 < CCC < 2$ des classes sont possibles, mais à vérifier. Le critère n'est pas applicable pour $CCC < 0$. Notons que ce critère n'est pas applicable si on a des variables corrélées, dans ce cas une analyse en composantes principales est envisageable.

Le pseudo F

Le pseudo F mesure la séparation entre toutes les classes, il est calculé comme [14] :

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}.$$

La valeur d'indice indiquant le nombre de classes à retenir doit être la plus grande possible, la valeur est choisie après le dernier « saut »important faisant augmenter l'indice en passant d'un nombre de classe $k-1$ à k.

Le pseudo t-carré

Le pic maximal de la valeur du pseudo t-carré indique le nombre $k-1$ de classes à retenir. Il faut donc ajouter un au nombre de classes correspondant à ce pic.

$$\text{pseudo } t^2 = \frac{B_{kl}}{((W_k + W_l)/(n_k + n_l - 2))},$$

où n_k et n_l sont le nombre d'observations des classes k et l, W_k et W_l sont la somme des carrés des classes k et l et B_{kl} la somme des carrés entre ces classes [15].

La méthode du coude

La méthode du coude [41] consiste à repérer un coude sur un graphe représentant, dans notre cas, la variable $1-R^2$ en fonction du nombre de classes. Le nombre de classes correspondant à ce coude est le nombre de classes à retenir. Sur l'exemple présenté en figure 5.1, le nombre de classes retenues est soit de 2 soit de 4, deux coudes sont présents.

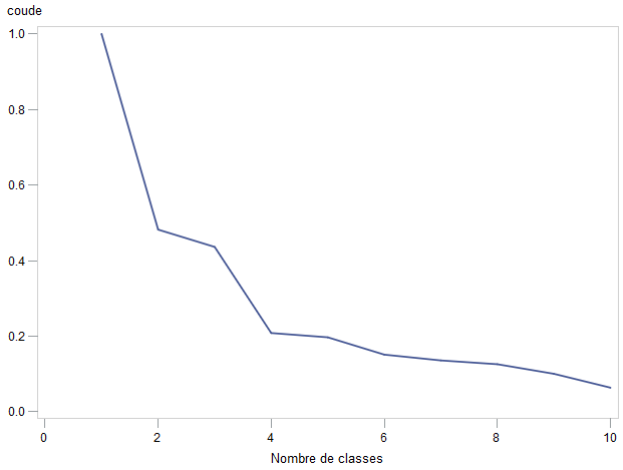


FIGURE 5.1 – Exemple pour la méthode du coude

Le dendrogramme

Les méthodes de classification hiérarchiques représentent l'ensemble des données par des partitions qui sont « emboitées ». En effet, deux classes ont soit une intersection vide, soit l'une est comprise dans l'autre, et la plus grande des classes contient tous les éléments à classer. La représentation graphique d'une hiérarchie est un dendrogramme. La hauteur des lignes regroupant les observations x_i dépend de la distance entre ces éléments (distance euclidienne par exemple). La figure 5.2 est un exemple de dendrogramme. Une droite horizontale grise est tracée au niveau où le dendrogramme fait le plus grand « saut », le nombre d'intersections entre cette droite et le dendrogramme désigne le nombre de classes à retenir. Dans l'exemple, le nombre de classes retenu est de 3, la première classe contient l'élément x_1, la deuxième les éléments x_4, x_5 et x_6 et la dernière les éléments x_2, x_7 et x_3.

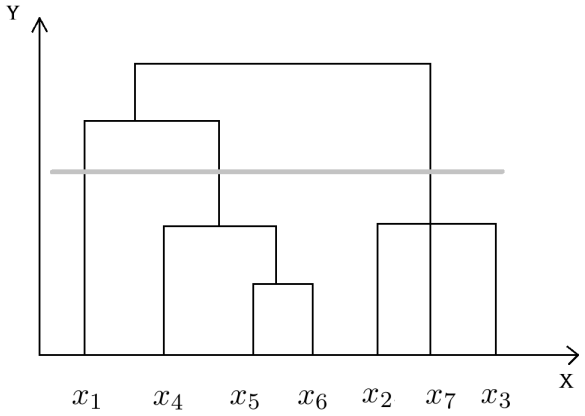


FIGURE 5.2 – Exemple de dendrogramme

Choix du nombre de classes

Afin de déterminer le nombre de classes à retenir, deux cas de figure peuvent se présenter. Dans le premier cas où les différentes méthodes de choix du nombre de classes donnent le même résultat, c'est ce nombre qui sera retenu. Dans le cas où différents nombres sont retenus par les méthodes, les différentes classifications doivent être analysées afin de retenir celle qui s'adapte le mieux aux données. Une étiquette est alors attribuée aux différentes classes créées.

2 Les méthodes de classification

Nous utilisons deux types de classifications : les méthodes hiérarchiques et les méthodes de partitionnement qui ont été évoquées précédemment. Pour les méthodes hiérarchiques, une nouvelle classe de la classification en k classes est créée à partir de l'union de deux classes de la classification en $k+1$ classes.

Nous utiliserons plusieurs méthodes de classification hiérarchiques : la méthode du lien minimum, du lien moyen, de Ward, du centroïde et une méthode de partitionnement : la méthode des nuées dynamiques.

L'algorithme des classifications ascendantes hiérarchiques consiste à former un nombre k de classes à partir des $k+1$ classes précédentes, en unissant deux de ces classes. Lors de la première étape de l'algorithme, il y a donc n classes (où n est le nombre d'éléments à classer), chaque classe contenant un seul élément. Deux classes sont regroupées si le couple formé par ces classes minimise l'indice d'agrégation δ. Si il y a plus d'un couple qui minimise cet indice, le premier couple rencontré est agrégé, on obtient alors une hiérarchie binaire.

La **méthode du lien minimum** rassemble les classes dont la distance minimale, entre ces classes (entre deux objets de ces classes), est plus petite que la distance minimale entre deux autres classes.

La **méthode du centroïde** regroupe à chaque étape les classes dont les centres de gravité sont les plus proches. Le centre de gravité d'une classe A est calculé comme

$$\frac{\sum_{x_i \in A} x_i}{n_A},$$

où n_A est le nombre d'éléments de la classe A et les x_i sont les observations (appartenant à la classe A).

La **méthode de Ward** consiste à regrouper deux classes de façon à ce que l'augmentation de l'intertie associée à la partition soit la plus faible possible. L'inertie d'un ensemble E par

rapport à un point a est définie comme

$$I_a(E) = \sum_{i=1}^{n} d^2(x_i, a),$$

où d est une distance euclidienne.

La **méthode de partitionnement des nuées dynamiques** consiste à simultanément chercher une partition en classes C_i et une représentation de cette partition en classes L_i. A chaque classe C_i est associée un *prototype* L_i. Le critère à optimiser est

$$\sum_{i=1}^{k} D(C_i, L_i),$$

où D mesure la dissemblance entre les individus et les prototypes des classes. Une décroissance de ce critère exprime une meilleure adéquation entre les classes recherchées et les prototypes associés.

Chapitre 6

Régression linéaire et modèle autorégressif

1 Régression linéaire

Une régression est une méthode statistique permettant d'analyser la relation entre une variable, appelée *variable dépendante*, par rapport à d'autres variables, appelées *variables explicatives* ou *régresseurs*.

Une régression linéaire fait l'hypothèse que la relation entre la variable dépendante y et les variables explicatives x_i est linéaire. De manière générale, le modèle linéaire s'écrit

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

où les β_i sont des paramètres à estimer, pour que la fonction de régression décrive le mieux les données et ϵ l'erreur. Pour cela, le système suivant est résolu

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \epsilon_i,$$

où p est le nombre de variables du modèle et les ϵ_i sont appelés les erreurs. Les indices i représentent les observations et les indices j les régresseurs du modèle. Une seconde hypothèse du modèle est que les erreurs suivent une loi normale et sont indépendantes. Notons que la régression linéaire consiste à résoudre un système d'équation surdéterminé [28], le nombre d'observations doit donc être plus important que le nombre de variables du modèle. Les notations suivantes sont utilisées :

$$\begin{aligned} SST &= \sum (y_i - \bar{y})^2; \\ SSE &= \sum (y_i - \hat{y}_i)^2; \\ SSR &= \sum (\hat{y}_i - \bar{y})^2 = \sum e_i^2; \end{aligned}$$

où les $\hat{y}_i$ sont l'évaluation des x_i par la régression. SST est la somme des carrés totale, SSE la somme des carrés expliquée et SSR la somme des carrés due à la régression. Ces valeurs sont liées par la relation

$$SSE + SSR = SST.$$

Le *coefficient de détermination* R^2 traduit la variance expliquée par un modèle de régression, il est calculé comme

$$R^2 = 1 - \frac{SSE}{SST}.$$

Nous avons que $0 \leq R^2 \leq 1$, plus le coefficient est proche de 1, plus l'équation de régression est adaptée pour décrire la distribution des points. Un R^2 proche de 0 indique une relation non

linéaire entre la variable dépendante et les régresseurs. Le R^2 augmente dès qu'une variable est ajoutée au modèle, deux modèles ne comportant pas le même nombre de variables ne peuvent donc pas être comparés par rapport à leur R^2. Dans ce cas, le *coefficient de détermination ajusté* R_a^2 sera utilisé, ce coefficient prend en compte l'ajout de variables au modèle et est donné par :

$$R_a^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1},$$

où n est le nombre d'observation et k le nombre de variables du modèle.

2 Autorégression

La régression linéaire « classique », présentée dans la section précédente est basée sur plusieurs hypothèses, l'une d'elle est que les erreurs sont indépendantes les unes des autres. Cependant, pour les séries temporelles, les résidus sont généralement corrélés à travers le temps. Il n'est donc pas possible d'appliquer une régression linéaire classique.

Les conséquences du non respect de l'indépendance des résidus sont les suivantes [26] : premièrement, les tests statistiques de signification des paramètres et les seuils de confiance pour les valeurs prédites ne sont pas correctes. Deuxièmement, l'estimation des coefficients de régression n'est pas aussi efficaces que si l'autocorrélation des variable avait été prise en compte. Enfin, étant donné que les résidus ne sont pas indépendants, ils ne peuvent pas être utilisés pour améliorer la prédiction des valeurs futures.

L'autorégression résout ce problème en augmentant le modèle de régression avec un modèle autorégressif pour les erreurs aléatoires prenant en compte l'autocorrélation des erreurs. Le modèle suivant est utilisé :

$$y_i = \beta_i \hat{x}_i + \nu_i; \tag{6.1}$$
$$\nu_i = -\phi_1\nu_1 - \phi_2\nu_2 - \cdots - -\phi_m\nu_m + \epsilon_i; \tag{6.2}$$

où les ν_i sont du bruit blanc et sont indépendants et identiquement distribués selon une loi normale centrée. L'autorégression calcule simultanément les coefficients β et les paramètres d'erreur du modèle autorégressif ϕ, en faisant cela l'autorégression corrige les estimations de la régression.

Le modèle autorégressif [23] est estimé par la méthode du maximum de vraisemblance. Cette méthode est délicate car la fonction de vraisemblance est très complexe dû à l'interdépendance des valeurs. La fonction de vraisemblance est à maximiser par rapport aux paramètres ϕ ce qui correspond à la minimisation des erreurs du modèle.

Pour exécuter cette fonction, la procédure `AUTOREG` du logiciel SAS est utilisée. Cette procédure renvoi différents indices permettant d'évaluer la qualité du modèle créé. Ces indices sont présentés ci-dessous.

Les coefficients R^2

L'autorégression renvoi deux coefficients de détermination R^2 [25]. Le premier, appelé R^2 total est calculé comme

$$R_{tot}^2 = 1 - \frac{SSE}{SST}.$$

Cet indice est une mesure de la façon dont la valeur suivante est prédite par le modèle.
Le deuxième, appelé R^2 de la régression, est calculé comme

$$R_{reg}^2 = 1 - \frac{TSSE}{TSST},$$

où $TSST$ est semblable à SST mais où les y_i utilisés sont transformés (voir equations 6.1 et 6.2). $TSSE$ est la somme des carrés des erreurs du nouveau modèle. Ce coefficient mesure

l'ajustement du modèle après transformation pour tenir compte des autocorrélations. Il s'agit du coefficient R^2 de la régression transformée.

Le critère d'information d'Akaike (AIC)

Le critère d'information d'Akaike [25] (en anglais Akaike information criterion ou AIC) est défini par

$$AIC = 2k - 2\ln(L_{max}),$$

où k est le nombre de paramètres à estimer du modèle et L_{max} le maximum de la fonction de vraisemblance du modèle. Le modèle choisi sera celui dont l'indice AIC est le plus petit. Ce critère fait un compromis entre la qualité de l'ajustement et la complexité du modèle (le nombre de régresseurs utilisés). Il pénalise donc les modèles comprenant un grand nombre de paramètres (et donc de régresseurs).
Cet indice permet donc de comparer différents modèles, par contre il ne dit rien sur la qualité absolue du modèle, contrairement au R^2 qu'on souhaite approcher de 1.

Le critère d'information bayésien (SBC)

Le critère d'information bayésien(en anglais Schwarz's Bayesian information criterion) [25] est défini par

$$SBC = -2\ln(L_max) + \ln(n)k,$$

où n est le nombre d'observations. Comme l'indice AIC, meilleur est le modèle, plus petit sera l'indice SBC [12].

La statistique de Durbin–Watson

Le test de Durbin-Watson [18] est un test statistique utilisé pour détecter la présence d'autocorrélation dans les résidus d'une régression. La statistique du test est donnée par

$$S_j = \frac{\sum_{i=j+1}^{n}(e_i - e_{i-j})^2}{\sum_{i=1}^{n} e_i^2},$$

où e_i est le résidu associé à l'observation i.

La valeur de S est comprise entre 0 et 4. Si $S < 2$, il y a une une corrélation positive de la série, si $S < 1$ cela signifie que des termes d'erreurs successifs ont des valeurs proches les unes des autres. Si $S > 2$, les erreurs successives sont, en moyenne, fort différents les uns des autres, c'est-à-dire corrélés négativement. Pour tester l'autocorrélation positive au niveau de confiance α, la statistique du test est comparée à une borne inférieure et supérieure (respectivement $S_{L,\alpha}$ et $S_{U,\alpha}$) :

$$\begin{cases} S < S_{L,\alpha} \text{ erreurs positivement corrélées;} \\ S > S_{U,\alpha} \text{ aucune preuve statistique que les erreurs sont autocorrélées positivement;} \\ S_{L,\alpha} < S < S_{U,\alpha} \text{ test non concluant.} \end{cases} \tag{6.3}$$

Le test, pour une corrélation négative au niveau de confiance α, utilise la statistique $4 - S$ et la compare aux valeurs critiques inférieure et supérieure (respectivement $S_{L,\alpha}$ et $S_{U,\alpha}$) :

$$\begin{cases} (4-S) < S_{L,\alpha} \text{ erreurs négativement corrélées;} \\ (4-S) > S_{U,\alpha} \text{ aucune preuve statistique que les erreurs sont autocorrélées négativement;} \\ S_{L,\alpha} < (4-S) < S_{U,\alpha} \text{ test non concluant.} \end{cases} \tag{6.4}$$

Les valeurs critiques inférieure et supérieure sont généralement disponibles dans une table enregistrée dans le programme utilisant le test.

Chapitre 7

Mesure de l'impact des bivalves exogènes sur la biomasse de phytoplancton

Dans ce chapitre, nous tentons de mesurer l'influence des bivalves invasifs sur le taux de chlorophylle a. Comme expliqué lors du chapitre 1, la chlorophylle a est un bon indicateur de la biomasse de phytoplancton présente en suspension dans l'eau. Les séries temporelles sont d'abord scindées en différentes périodes par rapport à l'apparition et l'installation des espèces exogènes de bivalves. Une régression linéaire est ensuite effectuée afin de mesurer l'importance de l'influence des bivalves invasifs.

1 Classification en différentes périodes

Afin de quantifier l'impact de l'arrivée des espèces exogènes de bivalves, nous souhaitons expliquer l'influence des différents paramètres physicochimiques et des différentes espèces de macroinvertébrés (regroupées en taxons) sur la chlorophylle a. Pour pouvoir différencier l'influence de ces espèces exogènes avant et après leur installation sur les différentes stations sillonnant la Meuse, la première chose qui est réalisée est une classification des données sur chaque station. Les méthodes de classification utilisées sont celles du lien moyen, du centroïde, de Ward et des nuées dynamiques. Les biologistes ont une idée, à partir des données récoltées le long de la Meuse, de l'année à partir de laquelle les espèces invasives sont arrivées sur chaque station, et de l'année à partir de laquelle ces espèces ont réellement été « bien »implantées. Ces espèces ne sont pas arrivées en même temps sur chaque station ; en effet elles arrivent de façon tardive en amont de la Meuse car elles la remontent doucement. Une analyse de classification nous permet d'identifier ces différentes périodes. La classification est effectuée sur deux variables, celle contenant l'année d'observation des données et celle concernant la population de bivalves dits invasifs. Les différentes stations étudiées sont celles de Liège, Hastière et Sassey-sur-Meuse.

Le commanditaire avait donné un exemple de classification pour la station de Sassey-sur-Meuse, sur base de l'observation de la population des bivalves. Celle-ci est la suivante :

1. 1998-2002 : avant l'apparition des espèces invasives,
2. 2003-2004 : pendant l'installation des espèces invasives,
3. 2005-2007 : période où les espèces invasives sont installées.

Les différents indices de détermination du nombre de classes rendus par les différentes méthodes appliquées sont les mêmes, leurs graphes sont présentés en figure 7.1. L'indice CCC étant plus petit que 0 aucune conclusion ne peut en être tirée, le pseudo F est lui aussi inutilisable. L'indice pseudo t-carré a son pic maximal en 1, le nombre de classes optimal est donc de $1 + 1 = 2$ classes. La méthode du coude représentée en figure 7.2 renseigne $2, 3$ ou 4 classes. Les dendrogrammes des différentes méthodes sont différents car la hauteur des lignes reliant les

observation dépend de la mesure distance utilisée, mais leur interprétation est la même : deux classes sont retenues.

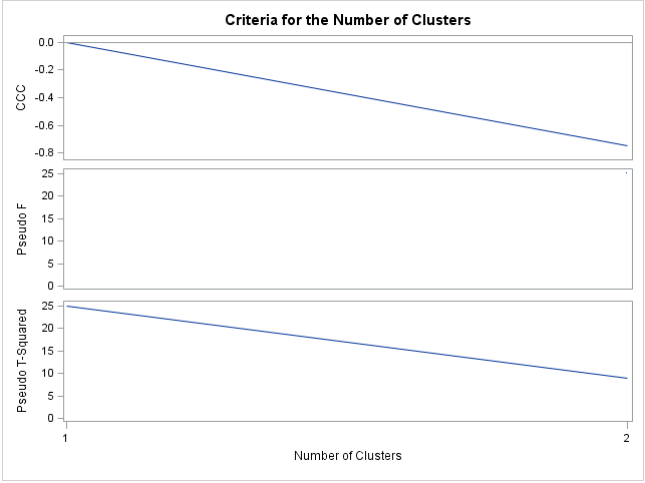


FIGURE 7.1 – Graphes des indices CCC, pseudo F et pseudo t^2 pour les classification effectuées sur la station de Sassey-sur-Meuse

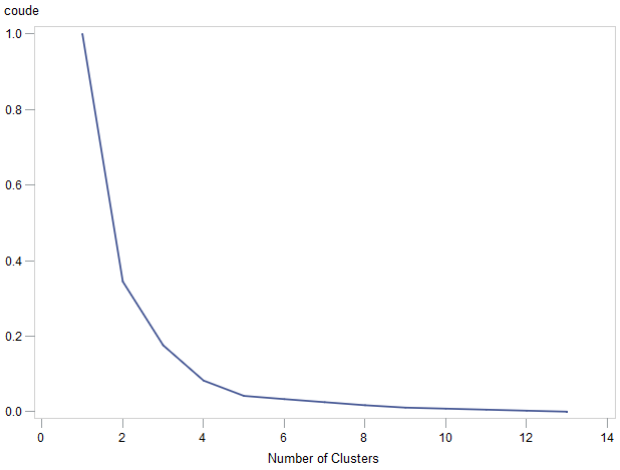


FIGURE 7.2 – Graphe de la méthode du coude pour la station de Sassey-sur-Meuse

Si 3 classes sont choisies, les classes proposées par le commanditaire sont retrouvées, celles-ci ne sont néanmoins pas prises en compte. En effet, le nombre d'observations total étant très petit (10) les classes comportent peu d'observations, ce qui ne se prête pas à notre étude car nous souhaitons réaliser une régression sur chaque classe. La classification en 4 classes est rejetée pour la même raison.

Les 2 classes retenues sont :

– 1998-2002 : avant l'apparition (significative) des espèces invasives,
– 2003-2007 : après l'apparition (significative) des espèces invasives.

La même étude est effectuée sur la station de Liège. Les différentes méthodes de classification rendent les mêmes graphes pour les indices CCC, pseudo F et pseudo t-carré, qui sont présentés en figure 7.3.

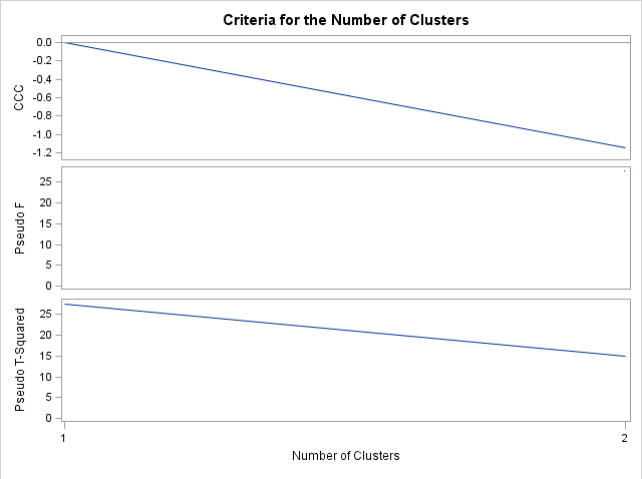


FIGURE 7.3 – Graphes des indices CCC, pseudo F et pseudo t^2 pour les classification effectuées sur la station de Liège

L'indice CCC (plus petit que 0) et pseudo F ne sont pas utilisés. Le pseudo t-carré indique qu'il faut retenir 2 classes, ce qui se confirme sur les différents dendrogrammes des méthodes de classifications appliquées et la méthode du coude. Les classes retenues sont :

- 1998-2002 : avant l'apparition (significative) des espèces invasives,
- 2003-2010 : après l'apparition (significative) des espèces invasives.

Pour la station de Hastière, seuls les dendrogrammes sont utilisables, ils indiquent un choix de 2 classes qui sont :

- 1998-2002 : avant l'apparition (significative) des espèces invasives,
- 2003-2005 : après l'apparition (significative) des espèces invasives.

La table 7.1 résume les différentes classifications réalisées précédemment en reprenant les périodes identifiées pour chaque station.

TABLE 7.1 – Différentes périodes détectées sur les stations étudiées

Station	période	observations	caractéristique
Sassey-sur-Meuse	1998-2002	5	avant l'apparition des espèces invasives de bivalves
	2003-2007	5	après l'apparition des espèces invasives de bivalves
Hastière	1998-2002	4	avant l'apparition des espèces invasives de bivalves
	2003-2005	4	après l'apparition des espèces invasives de bivalves
Liège	1998-2002	5	avant l'apparition des espèces invasives de bivalves
	2003-2010	5	après l'apparition des espèces invasives de bivalves

2 Choix du format des données

Nous disposons de 8 paramètres physicochimiques, d'une variable temporelle et de (maximum) 23 taxons de macroinvertébrés (certains taxons ne sont pas présents sur certaines stations) pour expliquer le taux de chlorophylle a sur les différentes stations. Le nombre d'observations par station varie de 8 à 10. Nous souhaitons appliquer une régression linéaire aux données afin de quantifier l'impact des bivalves exogènes sur la chlorophylle et de comparer leur influence avec celles des paramètres physicochimiques et des autres taxons.

La population de macroinvertébrés peut être exprimée de deux façons différentes : en nombre d'individus ou en pourcentage de la population totale (des macroinvertébrés). La population en nombre d'individus possède des valeurs qui varient fortement dans une même classe. Par exemple pour la classe des gastropoda, sur la station de Liège, la population varie d'environ 300 individus à plus 2000 individus certaines années. Or, en terme de pourcentage de population, une lourde chute d'individus ne signifie pas une grosse chute du pourcentage dans la population totale de macroinvertébrés. Par exemple en 1999 on dénombre 2296 gastropoda puis 689 individus en 2000, mais le pourcentage de gastropoda sur la population totale de macroinvertébrés varie relativement peu : il passe de 28.27% à 23.95%. Si les pourcentages peuvent lisser une hausse d'effectifs, ils peuvent aussi la cacher si une hausse d'individus d'une classe est accompagnée d'une baisse d'individus d'une autre classe, ainsi l'augmentation de la population de bivalves exogènes pourrait passer « inaperçue ». De plus, certains pourcentages sont très petits (de l'ordre de 10^{-4} et 10^{-5}) ce qui entraine que les coefficients des régressions sont très grands (ordre de 10^4 et 10^5).
En prenant les variables concernant les populations de macroinvertébrés sous forme de pourcentage de la population totale de macroinvertébrés, le coefficient R^2 est plus petit. Avec un R^2 et un nombre d'observations très petit, le coefficient R_a^2 s'avère être négatif. Nous choisissons d'utiliser les variables concernant les macroinvertébrés en nombre d'individus pour éviter ces inconvénients.

3 Diminution du nombre de variables

Afin de différencier l'influence des différents paramètres sur la chlorophylle a et l'impact de l'arrivée d'espèces exogènes sur les différentes stations, les observations sont divisées en plusieurs périodes. Le nombre d'observations qui était peu important (entre 8 et 10 par station) devient encore plus petit par rapport au nombre de variables. Il faut donc diminuer le nombre de variables afin de pouvoir appliquer une régression linéaire (le nombre de variables du modèle doit être inférieur au nombre d'observations).

3.1 Analyses en composantes principales

Pour diminuer le nombre de variables, les corrélations entre ces dernières sont observées. Si deux variables sont fortement corrélées (coefficient de corrélation de Pearson supérieur à 0.7 en valeur absolue), une des deux variables sera retirée du modèle. Si le nombre de variables est encore trop important, une analyse en composantes principales (acp) peut être effectuée. Cette analyse consiste à transformer des variables corrélées entre elles en variables décorrélées les unes des autres, appelées *composantes principales*. Cette analyse permet de réduire le nombre de variables. L'analyse en composantes principales à effectuer ne doit pas prendre en compte la chlorophylle a, variable à expliquer par la régression, ni les bivalves invasifs. En effet, si l'impact des bivalves invasifs sur la chlorophylle a veut être mesuré, il faut que cette variable soit isolée dans la régression et non agrégée dans une composante principale.

Deux choix sont possibles : le premier est d'effectuer une analyse en composantes principales sur toutes les variables non éliminées du modèle, autres que celles contenant le taux de chlorophylle a et la population de bivalves invasifs. Ensuite, une régression est appliquée, dont la variable indépendante est la chlorophylle a et les variables explicatives les bivalves invasifs

et la première composante principale de l'acp. Dans ce cas la droite de régression cherchée est de la forme :

$$chla = \beta_0 + \beta_{biv} * bivalvia\_invasifs + \beta_{CP} * CP,$$

où $bivalvia\_invasifs$ est la variable contenant la population de bivalves invasifs et CP est la première composante principale de l'acp réalisée sur toutes les variables restantes du modèle. Les β sont les coefficients réels calculés par la régression.

Le deuxième choix consiste à réaliser une analyse en composantes principales sur les paramètres physicochimiques (autres que *chla*) et une seconde acp sur les paramètres biologiques (autres que bivalves invasifs). La régression a pour variable indépendante la chlorophylle a et pour variables explicatives les bivalves invasifs ainsi que la première composante principale de chaque acp. Dans ce cas la droite de régression cherchée est de la forme :

$$chla = \beta_0 + \beta_{biv} * bivalvia\_invasifs + \beta_{physico} * CP_{physico} + \beta_{bio} * CP_{bio},$$

où $CP_{physico}$ et CP_{bio} sont respectivement la première composante principale des acp réalisées sur les paramètres physicochimiques et sur les paramètres biologiques. L'avantage de la deuxième méthode est de mieux cibler l'influence des paramètres sur la chlorophylle a.

Cette méthode ne semble pas judicieuse pour le commanditaire. D'une part l'acp rassemble des taxons qui évoluent de façon différente, d'autre part elle garde des taxons qui n'influencent pas la chlorophylle a. Une autre façon de diminuer le nombre de variables est donc abordée dans la section suivante.

3.2 Sélection des variables sur base de leur influence sur la chlorophylle a

Notre but est de diminuer le nombre de variables afin de pouvoir appliquer une régression linéaire expliquant le taux de chlorophylle a. La méthode présentée dans cette section se base l'importance de l'impact des variables sur le taux de chlorophylle a. Pour le commanditaire, les variables physicochimiques les plus importantes sont les paramètres dits « nutriments », car il s'agit de ce dont le phytoplancton se nourrit : l'ammonium, les nitrates et phosphates. Parmi les taxons de macroinvertébrés retenus, seuls les mollusques filtreurs ont une réelle influence sur la chlorophylle a, il s'agit des bivalves.

Une première régression expliquant le taux de chlorophylle a à partir des différents paramètres physicochimiques est effectuée. Celle-ci permet d'identifier les paramètres ayant le plus d'influence sur la chlorophylle. Une régression linéaire est ensuite appliquée aux données avec pour régresseurs les paramètres physicochimiques retenus et les populations de bivalves.

4 Étude station par station

4.1 Liège

Une première régression est effectuée sur les paramètres physicochimiques. En premier lieu, les corrélations entre ces variables, présentées en table 7.2, sont examinées. Une forte corrélation (0.91154) est observée entre les variables représentant le taux de phosphate et le taux de phosphore. La variable contenant le taux de phosphate PO_4^{3-} est retenue et celle concernant le phosphore $Ptot$ est retirée du modèle.

Deux types de régressions sont appliqués afin d'être comparés : une régression linéaire « classique »et une régression prenant en compte les autocorrélations. Les coefficients rendus par ces deux méthodes sont identiques et sont présentés sur la table 7.3. La régression linéaire classique possède un R^2 de 0.68 et un R^2 ajusté de 0.24 ce qui est très faible. Néanmoins, nous ne souhaitons pas comparer ce modèle à un autre, le R^2 suffit donc à nous indiquer la qualité du modèle. La régression prenant en compte les autocorrélations possède un R^2 total de 0.68. Les coefficients à minimiser sont $AIC = 78.86$ et $SBC = 83.38$.

TABLE 7.2 – Table des corrélations entre les différents paramètres physicochimiques à Liège

variable	mes	NH4+	NO3-	O2	PO43-	Ptot	Q	t
mes	1.00000	0.16552	-0.00256	0.11914	-0.05391	-0.09987	-0.04382	-0.37000
NH4+	0.16552	1.00000	0.16935	-0.08746	-0.11258	-0.21108	-0.16831	0.08904
NO3-	-0.00256	0.16935	1.00000	0.60181	0.16278	0.06472	-0.50505	0.07082
O2	0.11914	-0.08746	0.60181	1.00000	-0.07767	-0.23458	-0.32773	-0.01423
PO43-	-0.05391	-0.11258	0.16278	-0.07767	1.00000	0.91154	-0.45632	0.41284
Ptot	-0.09987	-0.21108	0.06472	-0.23458	0.91154	1.00000	-0.36329	0.36170
Q	-0.04382	-0.16831	-0.50505	-0.32773	-0.45632	-0.36329	1.00000	-0.59992
t	-0.37000	0.08904	0.07082	-0.01423	0.41284	0.36170	-0.59992	1.00000

TABLE 7.3 – Table des coefficients d'une régression linéaire effectuée sur les paramètres physicochimiques avec pour variable dépendante la chlorophylle a

variable	coefficient régression
	116.00956
mes	-0.84630
NH_4^+	-13.96439
NO^{3-}	-18.95793
O_2	0.66683
PO_4^{3-}	9.27923
Q	-0.01784
t	-2.13055

Le paramètre ayant le moins d'influence est le débit du cours d'eau Q. Ce paramètre n'ayant pas de tendance croissante ou décroissante (chapitre 4) il semble logique qu'il n'influence pas la décroissance de la chlorophylle a. Les paramètres les plus influents sont, comme prévus par le commanditaire, les nutriments du phytoplancton : l'ammonium, les phosphates et les nitrates.

Une régression est maintenant effectuée sur les paramètres physicochimiques retenus (ammonium, phosphates et nitrates) et les taxons influençant le taux de chlorophylle a : les bivalves natifs et invasifs. Les valeurs des coefficients β estimés est présentée sur la table 7.5. Le R^2 a baissé à 0.47 et le R^2 ajusté à 0.1029. Les coefficients AIC et SBC ont légèrement augmenté, ils valent maintenant respectivement 81.34 et 84.72.

TABLE 7.4 – Valeurs des coefficients estimés pour la régression sur les paramètres physicochimiques et biologiques à Liège

variable	coefficient
	26.6665
NH_4^+	6.7194
NO^{3-}	-9.4941
PO_4^{3-}	11.3614
bivalves invasifs	2.4485
bivalves natifs	-0.2239

L'impact des bivalves invasifs sur le taux de chlorophylle a est bien plus important que celui des bivalves natifs. Moins il y a de chlorophylle a, moins il y a de bivalves invasifs, mais plus il y a de bivalves natifs. La relation avec les paramètres physicochimiques n'est pas détaillée car elle n'est pas significative dans la mesure où de nombreux facteurs anthropiques affectent ces paramètres (notamment le taux de nitrates comme expliqué lors du chapitre 4).

La même procédure est appliquée sur les périodes décrites au chapitre 5, le but étant de voir si l'impact des bivalves invasifs change fortement entre ces deux périodes. Pour la deuxième période (2002-2010) nous obtenons

$$chla = -68.86 + 107.28 * NH_4^+ + 4.84 * NO^{3-} + 33.7 * PO_4^{3-} + 5.01 * \text{bivalves invasifs} + 0.45 * \text{bivalves natifs}.$$

Les paramètres repris dans le modèle auraient ainsi tous un impact positif sur le taux de chlorophylle a (plus leur valeur augmente, plus le taux de chlorophylle augmente). Le R^2 est de 0.8269 et le R^2 ajusté est de 0.7443 le modèle ajuste donc assez bien les données.

La première période (1998-2002) comporte 5 observations, il faut donc diminuer le nombre de variables. Pour cela, une analyse en composantes principales est effectuée sur les 3 variables physicochimiques. La première composante principale explique 39.51 % de la variance, elle est donc un bon candidat pour réduire les 3 paramètres physicochimiques en une seule variable. Nous choisirons néanmoins la deuxième composante, notée $CP_{physico}$, qui répartit mieux les variables sur son axe, comme présenté en figure 7.4. Ainsi des valeurs négatives de la deuxième composante principale, se rapprochent de l'ammonium, des valeurs positives du phosphate et de petites valeurs du nitrate. Cette composante conserve 37.08 % de la variance, elle reste donc très correcte par rapport à la première composante principale. La variable $CP_{physico}$ est définie par le vecteur propre : (0.731985 PO_4^{3-}, 0.034931 NO^{3-}, -0.680424 NH_4^+).

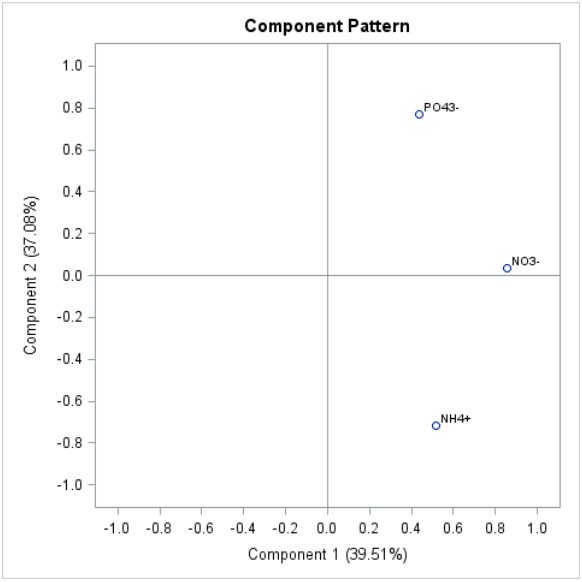


FIGURE 7.4 – Représentation des deux premières composantes principales de l'acp réalisée sur les paramètres physicochimiques du modèle de régression de Liège

L'estimation des paramètres des régressions effectuées est présentée dans la table 7.5. Selon ces régressions linéaires, les bivalves invasifs ont une relation positive avec le taux de chlorophylle a, tandis que les paramètres physicochimiques nutriments ont une relation négative avec ce taux. Notons également que d'une période à l'autre, l'importance des bivalves invasifs augmente, et la relation entre la chlorophylle a et les bivalves natifs s'inverse.

En effet, sans tirer de conclusions trop hâtives, et sans parler de cause à effet, il semble logique que plus il y a de chlorophylle a, plus il y a de bivalves car les bivalves se nourrissent de phytoplancton dont la chlorophylle a est un bon marqueur de la biomasse. Lors de la deuxième période, où les bivalves invasifs sont installés à Liège, la tendance s'inverse pour les bivalves autochtones, plus il y a de chlorophylle a, moins il y a de ces bivalves, alors que les bivalves invasifs renforcent leur position (plus il y a de chlorophylle a, plus il y en a). On pourrait donc interpréter ceci par le fait que les bivalves invasifs « profitent »plus de la chlorophylle a (du phytoplancton) au détriment des bivalves « belges ».

TABLE 7.5 – Coefficients des régressions expliquant la chlorophylle a effectuées sur les deux périodes temporelles à Liège

variable	coefficient période 1	coefficient période 2
	0.47682	-10.08503
$CP_{physico}$	-0.63756	-1.18669
bivalves invasifs	1.22326	7.55466
bivalves natifs	4.73596	-1.89514
R^2	0.9042	0.8666
R^2_a	0.6167	0.7665

4.2 Hastière

Comme pour la station de Liège, nous commençons par une première régression linéaire sur les paramètres physicochimiques afin de ne retenir que ceux ayant une influence notable sur la chlorophylle a. Il ressort de la première analyse sur les corrélations, que les variables *Ptot* et PO_4^{3-} sont fortement corrélées (0.90455) ainsi que les variables O_2 et NH_4^+ (-0.81306). Les variables PO_4^{3-} et NH_4^+ sont conservées dans le modèle tandis que les deux autres en sont enlevées.

TABLE 7.6 – Coefficients de la régression expliquant la chlorophylle a à partir des autres paramètres physicochimiques sur la station de Hastière

variable	coefficient
	65.67718
mes	-0.03982
NH_4^+	27.52306
NO^{3-}	-10.45545
PO_4^{3-}	92.03752
Q	0.03963
t	-3.17498

Le modèle possède un R^2 de 0.9175 et un R^2 ajusté de 0.4228. Étant donné que nous ne souhaitons pas comparer ce modèle à un autre modèle possédant un nombre différent de régresseurs, le R^2 suffit pour affirmer que le modèle ajuste bien les données. Le coefficient AIC est de 33.4948804, ce qui est bien meilleur que la régression similaire réalisée à Liège, il en va de même pour le coefficient SBC valant 34.05. Les variables ayant le plus d'influence sont les paramètres nutriments NH_4^+, NO^{3-} et PO_4^{3-}. L'influence des autres paramètres étant négligeable par rapport aux 3 nutriments, ils ne sont pas pris en compte dans le modèle expliquant la chlorophylle a. Comme pour Liège, une acp est effectuée sur les paramètres physicochimiques conservés afin de diminuer le nombre de variables pour effectuer une régression sur les deux périodes de temps : avant et après l'apparition significative des bivalves. Les deux premières composantes principales sont représentées en figure 7.5.

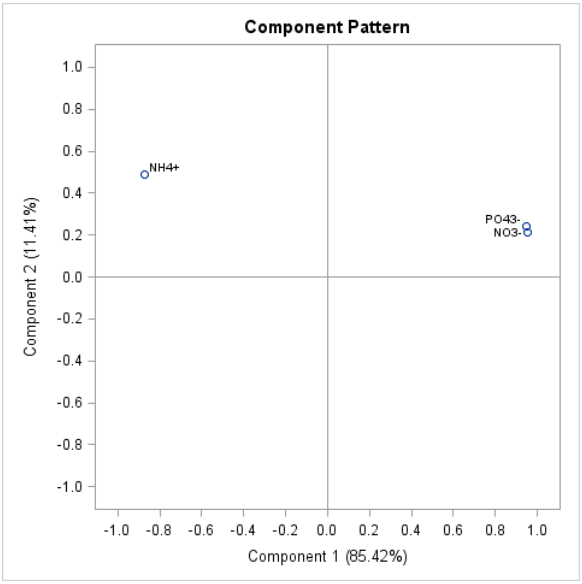


FIGURE 7.5 – Représentation des deux premières composantes principales de l'acp réalisée sur les paramètres physicochimiques du modèle de régression d'Hastière

La première composante principale sera utilisée pour représenter les paramètres physicochimiques retenus. Cette composante explique 85.42 % de la variance, remarquons que, comme pour la composante principale retenue pour la station de Liège, une valeur négative de la composante signifiera un plus grand taux d'ammonium. Cette composante est définie par le vecteur propre(0.591494 PO_4^{3-}, 0.594790 NO^{3-}, -0.544389 NH_4^+).

Une régression est appliquée à chaque période définie sur la station d'Hastière, les valeurs estimées des coefficients des régressions sont présentées sur la table 7.7. La deuxième période ne comportant que 3 observations, la composante principale retenue peut être écrite comme une combinaison linéaire des autres variables : $CP_{physico} = -8.66387 - 5.38256 * Bivalvianatifs + 9.60868 * Bivalviainvasifs$ et sa valeur est fixée à 0. L'influence des autres paramètres physicochimiques n'a que peu d'influence sur le taux de chlorophylle a par rapport à l'influence des bivalves. Durant la première période, où les bivalves invasifs sont peu nombreux, parfois inexistants (0 individus mesurés jusqu'en 2000) leur influence est moindre par rapport à celle des bivalves natifs. L'influence des bivalves est négative, cela signifie que plus il y a de bivalves, moins il y a de chlorophylle a. Lors de la deuxième période, le coefficient des bivalves natifs change de signe, par conséquent, plus il y a des chlorophylle a, plus il y a de bivalves natifs. Inversement moins il y a de chlorophylle a, moins il y a de bivalves natifs ; c'est cette interprétation qu'il faut privilégier car comme démontré durant le chapitre 4 le taux de chlorophylle a tendance à décroitre au fil des années. Par contre, les bivalves invasifs conservent le signe négatif de leur coefficient, donc moins il y a de chlorophylle a, plus il y a de bivalves invasifs.

TABLE 7.7 – Coefficients des régressions expliquant la chlorophylle a effectuées sur les deux périodes temporelles à Hastière

variable	coefficient période 1	coefficient période 2
	275.17951	27.95930
$CP_{physico}$	-9.83400	0
bivalves invasifs	-83.57060	-18.56281
bivalves natifs	-307.92996	9.35044
R^2	0.8854	1
R_a^2	0.5417	.

4.3 Sassey-sur-Meuse

Toujours dans la même optique, nous observons les corrélations entre les paramètres physicochimiques de Sassey-sur-Meuse. Une forte corrélation est observée entre les variables NH_4^+ et t (0.90269), la variable t est donc retirée du modèle. Les estimations de la régression effectuée sur les paramètres physicochimiques afin d'expliquer la chlorophylle a sont présentées dans la table 7.8. Le coefficient de détermination R^2 vaut 0.8872, la régression ajuste donc bien les données.

Contrairement aux stations précédemment étudiées, les nitrates jouent un rôle moindre par rapport aux autres variables à Sassey-sur-Meuse. Le phosphore généralement fortement corrélé aux nitrates ne l'est que très faiblement (0.26516) et son impact sur la taux de chlorophylle a est notable. Les variables retenues pour le modèle sont NH_4^+ ,PO_4^{3-} et $Ptot$. L'impact de l'oxygène est moindre par rapport à celui des variables retenues, de plus il possède une corrélation (0.77829) avec les variables $Ptot$, pour ces raisons le taux d'oxygène ne sera pas inclus dans le modèle incluant les paramètres biologiques.

TABLE 7.8 – Coefficients de la régression expliquant la chlorophylle a à partir des autres paramètres physicochimiques sur la station de Sassey

variable	coefficient
	-120.36429
mes	-1.92014
NH_4^+	201.34829
NO^{3-}	-0.70757
O_2	15.83080
PO_4^{3-}	-819.86897
$Ptot$	-214.63704
Q	0.82767

Une analyse en composantes principales est effectuée sur les paramètres physicochimiques retenus, les deux premières composantes principales sont représentées en figure 7.6. La première composante principale est utilisée dans la régression car elle conserve le plus d'information à propos des données. Elle est définie par le vecteur propre (0.620012 NH_4^+, 0.642362 PO_4^{3-}, 0.450507 $Ptot$).

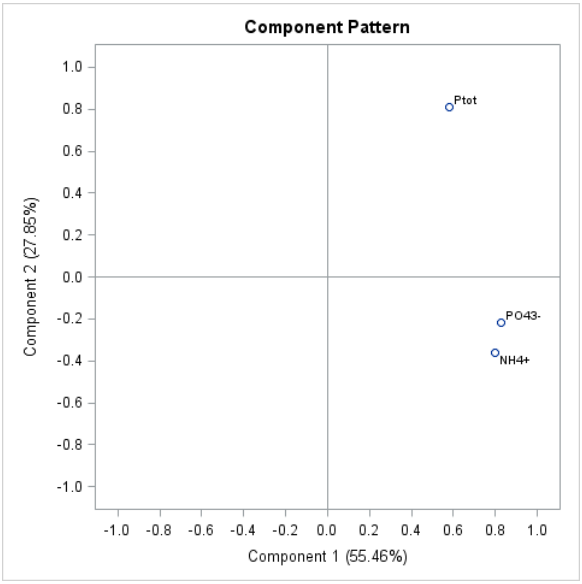


FIGURE 7.6 – Représentation des deux premières composantes principales de l'acp réalisée sur les paramètres physicochimiques du modèle de régression de Sassey

Les résultats des régressions réalisées sur chaque période sont représentés dans la table 7.9. La première période est marquée par une faible importance des paramètres physicochimiques qui évoluent de façon similaire à la chlorophylle a, tandis que les bivalves ont un effet presque similaire (légèrement plus fort pour les bivalves natifs) mais de signe opposé. Moins il y a de chlorophylle a, plus nombre de bivalves invasifs diminue et plus celui des bivalves natifs augmente. Lors de la deuxième période, où les bivalves invasifs sont implantés, l'impact des autres paramètres physicochimiques augmente et change de signe. Ainsi, moins il y a de chlorophylle a, plus il y a d'ammonium, de phosphates et de phosphore. Nous pourrions conclure que le taux de nutriments reste le même mais que ceux-ci sont moins absorbés car la biomasse de phytoplancton diminue. Or, comme nous l'avons précisé dans le chapitre sur les tendances, de nombreux facteurs anthropiques interviennent quant à l'apport, entre autre, de nitrates dans la Meuse, les liens entre ces variables sont alors trop compliqués à expliquer par notre modèle

car ils recèlent de nombreux effets « cachés ». L'impact respectif des bivalves reste de même signe mais s'accentue fortement sur la deuxième période.

TABLE 7.9 – Table des corrélations entre les différents paramètres physicochimiques à Sassey

variable	coefficient période 1	coefficient période 2
	29.28462	-95.01180
$CP_{physico}$	2.07210	-7.82473
bivalves invasifs	10.62469	173.47111
bivalves natifs	-13.88562	-99.11360
R^2	0.9850	0.6957
R_a^2	0.9402	-0.2170

5 Conclusion

D'une part nous avons les stations de Sassey-sur-Meuse et Hastière où la population des bivalves invasifs est très faible avant 2002 (période 1), d'autre part la population de ces bivalves à Liège est, certes, moins abondante avant 2002, mais déjà plus imposante que sur les deux premières stations citées. Avec le temps, les bivalves invasifs ont tendance à avoir une relation positive avec la chlorophylle a : c'est-à-dire que moins il y a de chlorophylle a, plus il y a de bivalves. Aux stations des Liège et Sassey où le coefficient de régression de ces bivalves est positif, celui ci augmente de la première à la deuxième période, la relation décrite précedemment s'accentue donc. Sur la station d'Hastière, où le coefficient est négatif en première période, il augmente (même si il reste négatif on peut imaginer qu'avec une période de temps plus longue celui-ci devienne positif). On peut donc observer un réel impact : plus il y a de bivalves exogènes, moins il y a de chlorophylle a. Aucune conclusion de cause à effet ne peut néanmoins être faite. En effet, des facteurs intermédiaires non pris en compte pourraient intervenir.

Pour les bivalves natifs, nous observons deux comportements différents : d'une part à Liège et Sassey les bivalves natifs tendent à avoir un coefficient à signe de plus en plus négatifs, ainsi moins il y a de chlorophylle a, plus il y a de bivalves natifs ; d'autre part à Hastière ce coefficient tend à devenir de plus en plus grand, ainsi moins il y a de chlorophylle a, moins il y a de bivalves natifs. La majeure différence entre ces deux cas est que le signe du coefficient concernant les bivalves invasifs est négatif à Hastière.
Ainsi, si la relation bivalves invasifs - chlorophylle a est négative, celle entre les bivalves natifs serait plutôt positive.

Remarquons finalement que l'influence des deux types de bivalves est généralement opposée lors de la deuxième période. Cela signifie que si l'une des population augmente, l'autre diminue, et inversement.

Chapitre 8

Analyse de Co-inertie

L'analyse de co-inertie est une méthode de couplage entre deux tableaux. Il est donc possible de l'appliquer à deux jeux de données qui ont quelque chose en commun. En général (et pour plus de facilité dans le logiciel R) on suppose que la variable en commun est en ligne sur les tableaux. La co-inertie est appliquée à nos données avec la variable temporelle en commun.

Il s'agit d'une méthode très souvent utilisée en biologie, une simple recherche sur le net permet de se rendre compte du nombre d'articles basés sur cette méthode dont les exemples sont donnés avec des données biologiques, notamment [35],[48] et [51].

Afin de coupler nos tableaux de données, une analyse en composantes principales est tout d'abord réalisée sur les deux tableaux de l'étude.

Deux BCA (Between-Class Analysis) sont ensuite effectuées entre chaque acp et la variable reprenant les dates d'observation des données. Ces analyses permettent d'insister sur le caractère temporel des données. Une analyse de co-inertie est enfin réalisée sur les deux BCA précédemment construites.

Nous pouvons ainsi représenter sur un même plan de dimension 2 les dates d'observation des données, les paramètres physicochimiques et les taxons de macroinvertébrés. Cette représentation permet d'analyser le comportement des variables les unes par rapport aux autres et leur comportement à travers le temps.

N.B. : dans ce chapitre, les taxons sont exprimés en pourcentage de la population totale de macroinvertébrés. Des analyses similaires peuvent être effectuées avec les taxons exprimés en nombre d'individus. Les résultats donnent des interprétations similaires. Les évolutions des bivalves et de la chlorophylle a que nous avons commentées restent les mêmes.

1 Analyse de la station de Liège

Dans un premier temps, une analyse en composantes principales est appliquée à chaque table de données. Les variances expliquées par chacune des composantes principales sont représentées en figure 8.1 pour les macroinvertébrés et 8.2 pour les paramètres physicochimiques. Seules les deux premières composantes principales sont conservées.

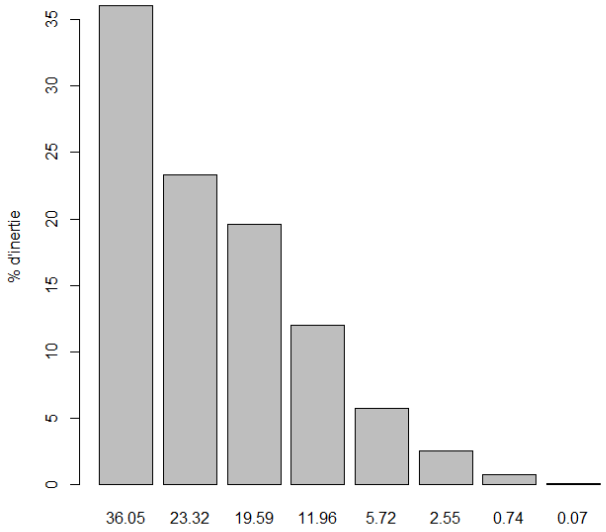


FIGURE 8.1 – Variance expliquée des différentes composantes principales pour l'acp sur les macroinvertébrés

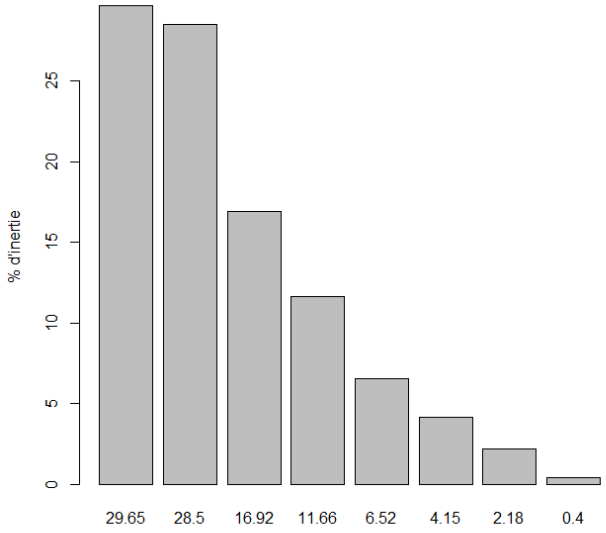


FIGURE 8.2 – Variance expliquée des différentes composantes principales pour l'acp sur les paramètres physicochimiques

Les résultats de l'analyse de co-inertie sont présentés en figure 8.8. Observons tout d'abord la répartition des dates sur la figure 8.3, ce graphe se lit en « cercle », nous partons du centre en 1998 pour se diriger vers le haut à droite au début des années 2000 et descendre jusqu'en 2006 pour enfin remonter vers la fin de la première décennie des années 2000.

L'origine du graphe $(0, 0)$ représente la situation moyenne de toutes les données projetées sur les deux composantes principales retenues. Nous observons que les observations s'éloignent peu à peu, au cours du temps, de la situation moyenne, dans des directions opposées, avec des pics d'éloignement en $2001, 2006$ et 2010. Remarquons que les années 2000 à 2002 semblent suivre un comportement atypique, comme si à cette époque là, un schéma se cherchait.

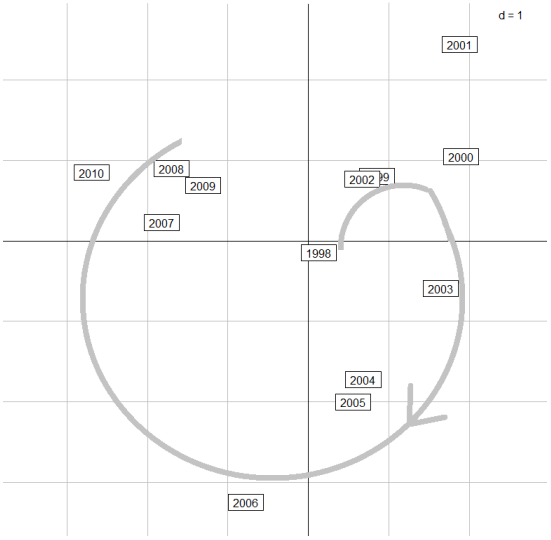


FIGURE 8.3 – Répartition des dates d'observation sur un plan à 2 dimensions

Une BCA (Between-Class Analysis) est effectuée entre chaque acp et la variable reprenant les dates d'observation des données. La BCA permet d'insister sur le caractère temporel des données. Ainsi les graphes 8.6 et 8.4 représentent respectivement les macroinvertébrés et les paramètres physicochimiques.

Sur la figure 8.4 nous pouvons observer les paramètres physicochimiques par rapport aux dates d'observation. Nous remarquons que la chlorophylle a *Chla* se situe sur le centre à droite du graphe, près de l'état « moyen »de 1998. Si nous reprenons la trajectoire temporelle de la figure 8.3 et que nous la traduisons en *Chla*, via les projections sur l'axe de la chlorophylle a, nous observons que cette dernière croît jusqu'en 2001, puis suit une décroissance marquée jusqu'en 2006 avec quasi stabilisation entre 2006 et 2010. Cette analyse décrit bien ce qui est visible sur la figure 8.5 représentant l'évolution de la chlorophylle a au cours du temps.

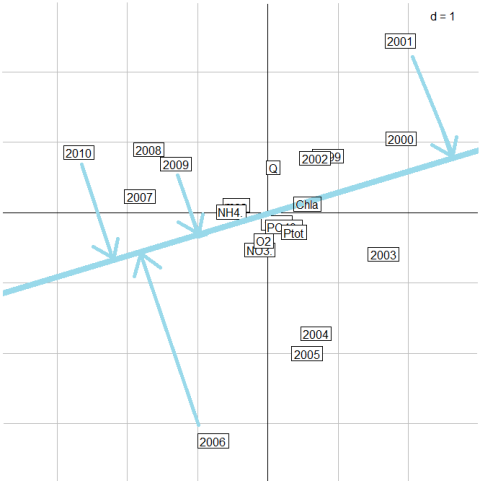


FIGURE 8.4 – Co-inertie entre les paramètres physicochimiques et les dates

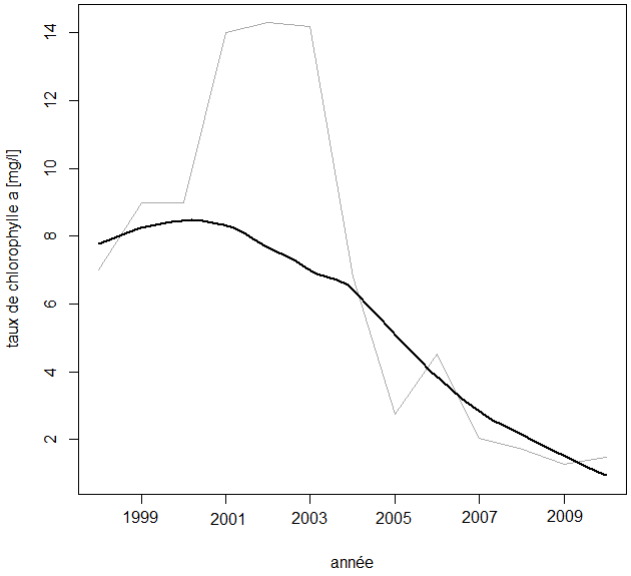


FIGURE 8.5 – Evolution de la chlorophylle a entre 1998 et 2010 à Liège en gris et courbe lowess en noir

En ce qui concerne les macroinvertébrés, la figure 8.6 ne permet pas une visualisation précise car les macroinvertébrés sont regroupés les uns sur les autres, nous nous baserons donc sur la figure agrandie (zoom) 8.7 pour notre analyse.

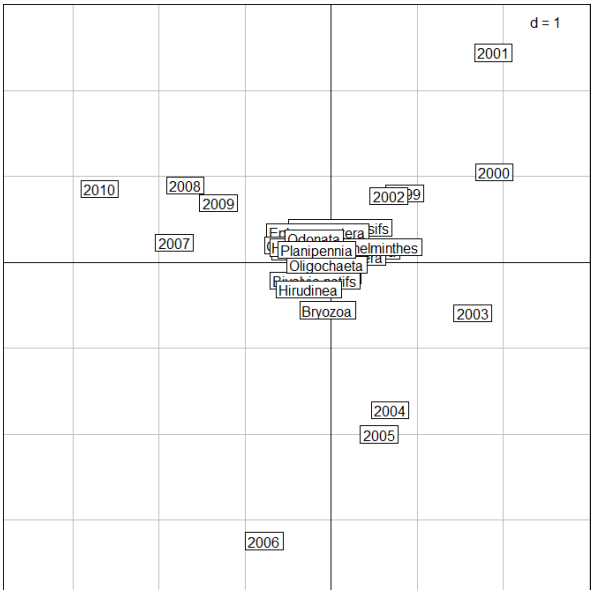


FIGURE 8.6 – Co-inertie entre les macroinvertébrés et les dates

Nous pouvons déjà observer que les comportements temporels des bivalves natifs et invasifs sont diamétralement opposés : plus la présence des uns augmente, plus les autres disparaissent. La présence de bivalves invasifs semble maximale en 2001, pour décroître jusqu'en 2006 et se stabiliser à un taux proche de la moyenne de 2010. Des analyses semblables peuvent être effectuées avec les autres taxons.

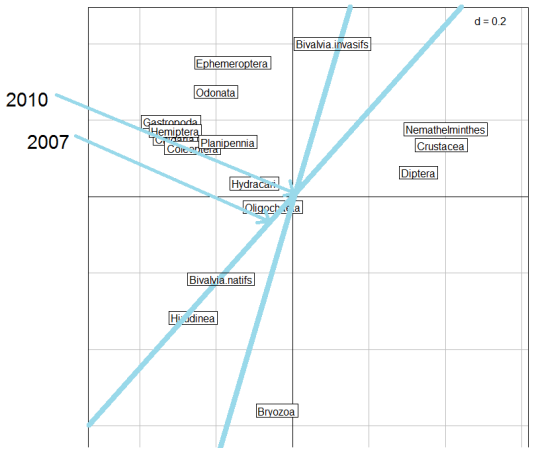


FIGURE 8.7 – Répartition des macroinvertébrés sur le plan

Le résumé de l'analyse est présenté en figure 8.8. Les flèches présentes derrière les dates représentent le lien entre les variables physicochimiques et les taxons de macroinvertébrés, plus la flèche est grande, moins la corrélation entre ces deux types de variables est grande. L'année 2007 semble une année de transition lorsque les deux types de variables sont très peu corrélées. Nous retrouvons également les axes liés aux variables physicochimiques et aux macroinvertébrés.

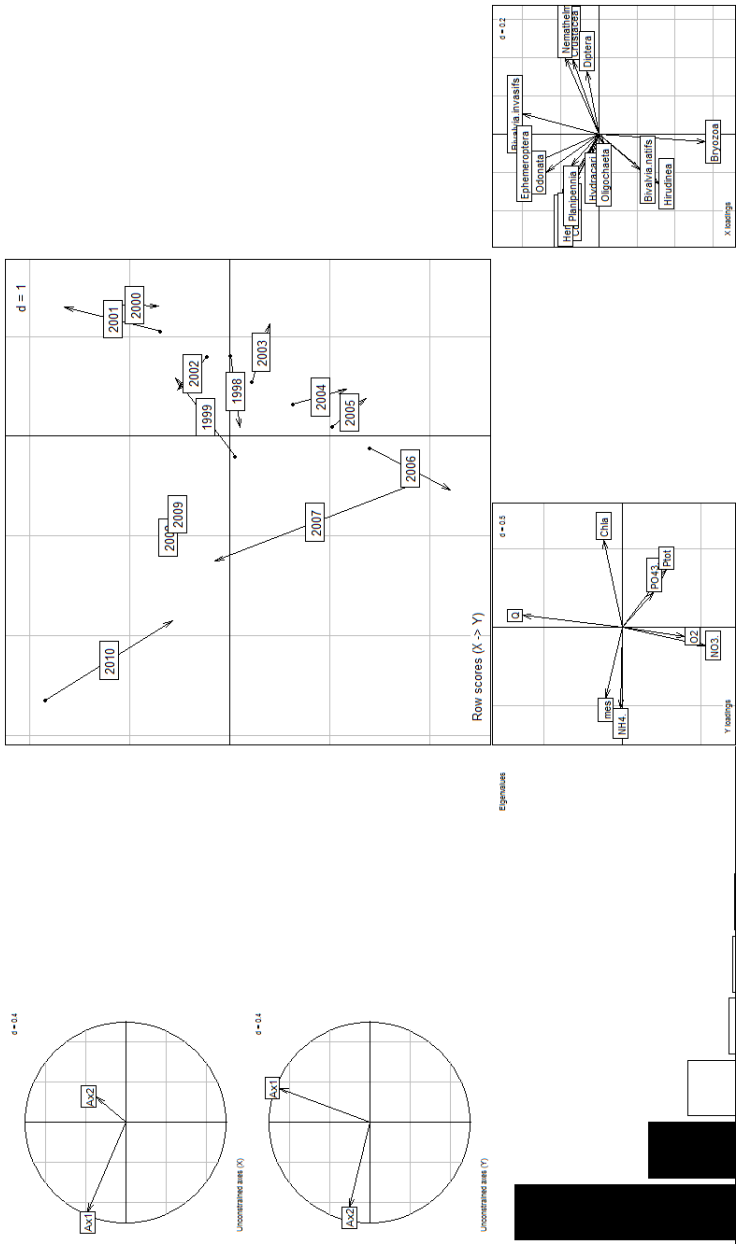


FIGURE 8.8 – Résultats de la co-inertie entre les paramètres physicochimiques et les macroinvertébrés à Liège

2 Analyse de la station de Hastière

L'analyse effectuée sur Liège est réalisée de la même manière à Hastière. Comme précédemment, nous disposons de deux tableaux, l'un contenant les populations des différents taxons de macroinvertébrés, l'autre contenant les mesures des paramètres physicochimiques de l'eau pour les années de 1998 à 2005.

Nous réalisons une acp sur chacun des tableaux. Les figures 8.9 et 8.10 représentent les variances expliquées de chacune des composantes principales des acp, respectivement réalisées sur les macroinvertébrés et sur les paramètres physicochimiques.

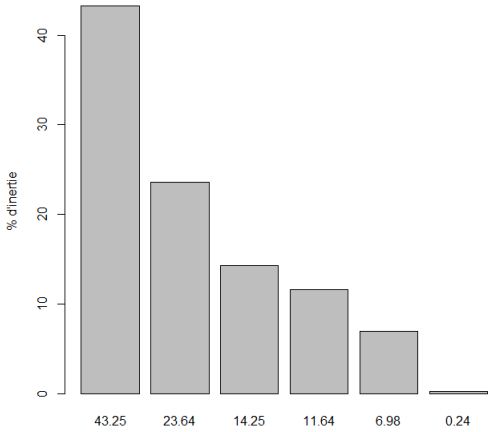


FIGURE 8.9 – Variance expliquée des différentes composantes principales pour l'acp sur les macroinvertébrés

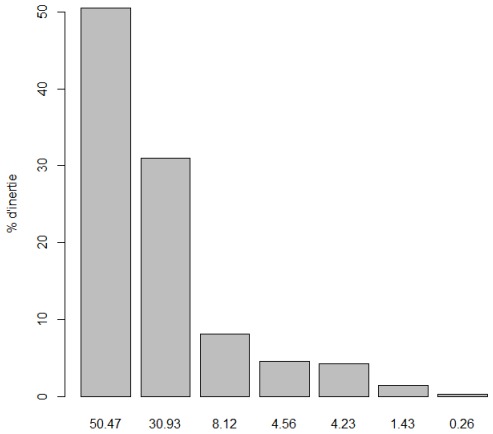


FIGURE 8.10 – Variance expliquée des différentes composantes principales pour l'acp sur les paramètres physicochimiques

La figure 8.11 présente la répartition des dates d'observation sur un plan en 2 dimensions. L'origine du graphe représente la situation moyenne de toutes les données projetées sur les deux composantes principales retenues. La situation moyenne se situe en 1999, les observations s'éloignent peu à peu, au cours du temps, de la situation moyenne, contrairement à Liège, la première date dont nous disposons (1998) est éloignée de la moyenne.

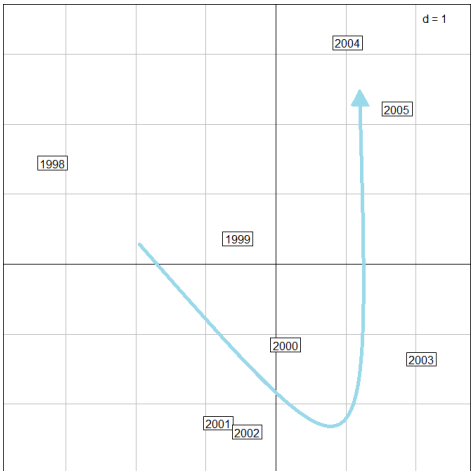


FIGURE 8.11 – Répartition des dates d'observation sur un plan à 2 dimensions

Sur la figure 8.12 les paramètres physicochimiques sont représentés par rapport aux dates d'observation. La chlorophylle a *Chla* est proche de l'état « moyen »de 1999. En projetant les différentes dates sur l'axe de la *Chla*, nous observons que cette dernière décroît jusqu'en 2000 et se stabilise ensuite durant 2 ans en 2001 et 2002 pour croître à partir de 2002 jusqu'à stabilisation entre 2004 et 2005.

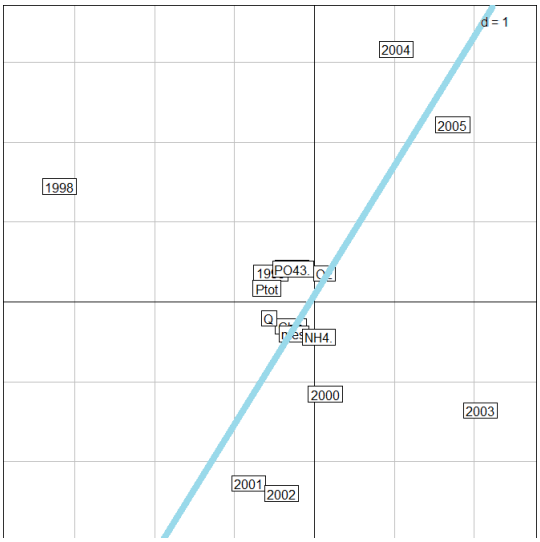


FIGURE 8.12 – Co-inertie entre les paramètres physicochimiques et les dates

Les bivalves invasifs et natifs ne sont pas diamétralement opposés comme à Liège. Après un début très en retrait de la moyenne en 1998, les bivalves natifs ou invasifs semblent ne pas évoluer fortement à partir de 2000 (sauf en 2003 où ils sont plus présents).

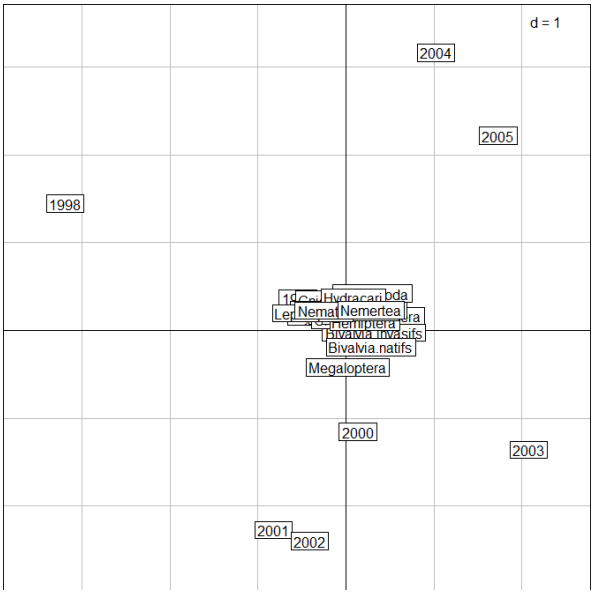


FIGURE 8.13 – Co-inertie entre les macroinvertébrés et les dates

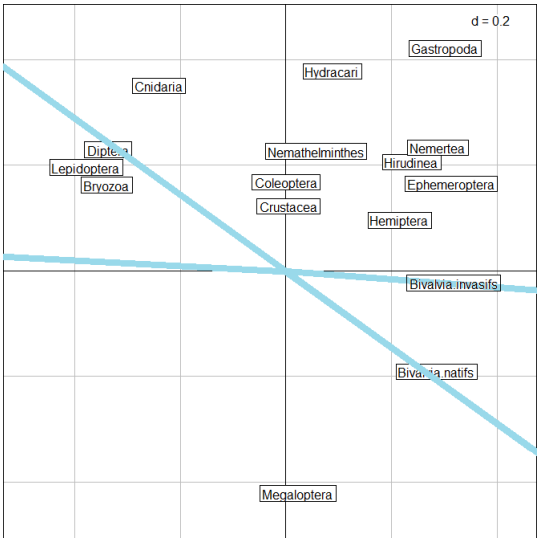


FIGURE 8.14 – Répartition des macroinvertébrés sur le plan

La figure 8.15 présente un résumé de la co-inertie.

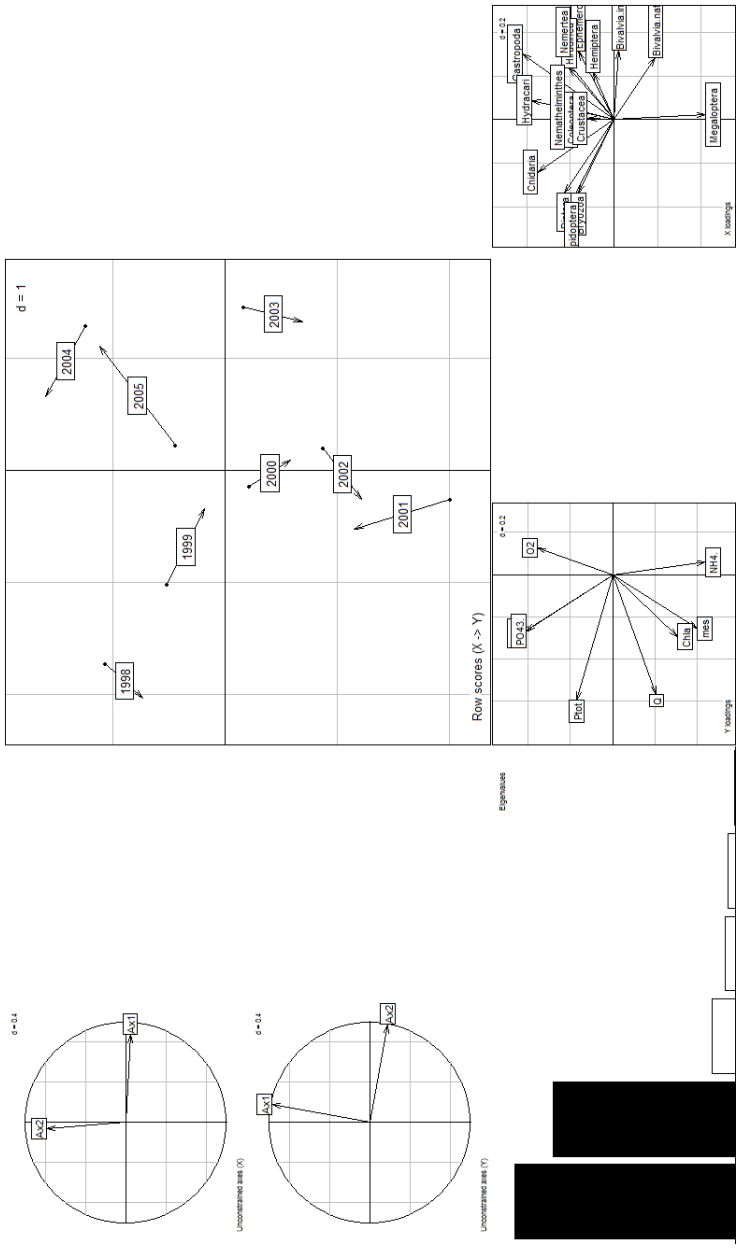


FIGURE 8.15 – Résultats de la co-inertie entre les paramètres physicochimiques et les macroinvertébrés à Hastiere

3 Analyse de la station de Sassey-sur-Meuse

Pour cette dernière station, nous disposons toujours de deux tableaux, l'un contenant les populations des différents taxons de macroinvertébrés, l'autre contenant les mesures des paramètres physicochimiques de l'eau pour les années de 1998 à 2007. Une analyse en composantes principales est réalisée sur chacun de ces tableaux. Les figures 8.9 et 8.10 représentent les variances expliquées de chacune des composantes principales des acp, respectivement réalisées sur les macroinvertébrés et sur les paramètres physicochimiques.

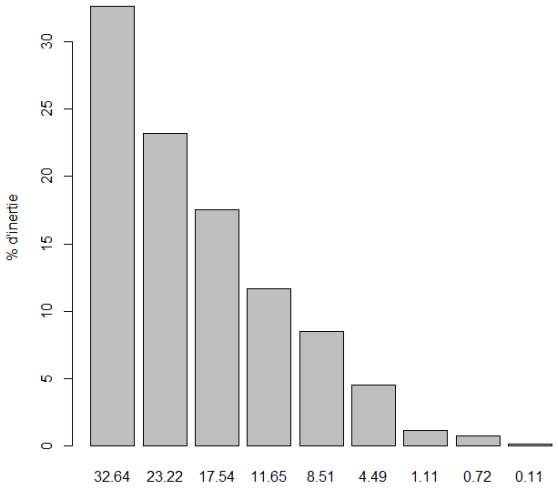


FIGURE 8.16 – Variance expliquée des différentes composantes principales pour l'acp sur les macroinvertébrés

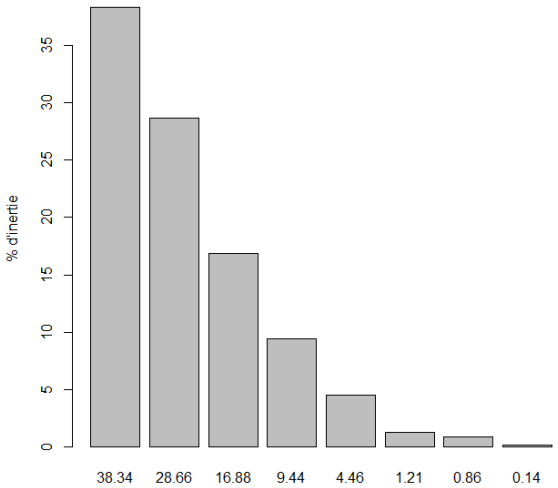


FIGURE 8.17 – Variance expliquée des différentes composantes principales pour l'acp sur les paramètres physicochimiques

Contrairement aux analyses précédents, la situation moyenne des données ne se situe pas

au début de la série temporelle mais plutôt vers la fin en 2004-2005.

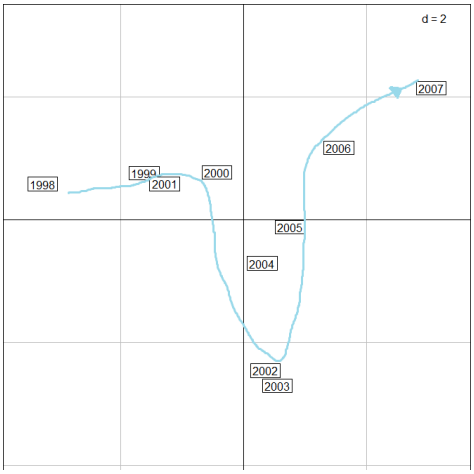


FIGURE 8.18 – Répartition des dates d'observation sur un plan à 2 dimensions

L'axe de la chlorophylle a est presque l'axe vertical du plan en 2 dimensions. Ainsi nous pouvons observer, après une situation stable entre 1998 et 2001, une augmentation important de la *Chla* en 2002 et 2003. Un retour vers des valeurs de départ s'observe à partir de 2004, la *Chla* diminuant fortement jusqu'en 2007.

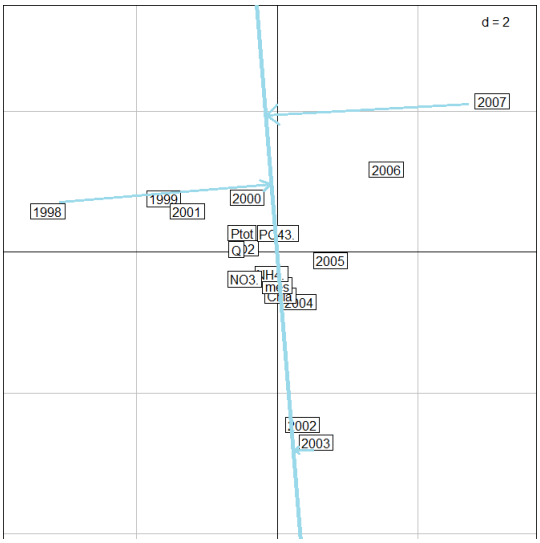


FIGURE 8.19 – Co-inertie entre les paramètres physicochimiques et les dates

Pour ce qui est des bivalves invasifs, nous observons une légère décroissance jusqu'en 2001, ensuite ceux-ci connaissent une croissance marquée jusqu'en 2007. Les bivalves natifs sont eux relativement stables jusqu'en 2002, décroissent en 2002-2003 pour enfin connaitre une période de croissance.

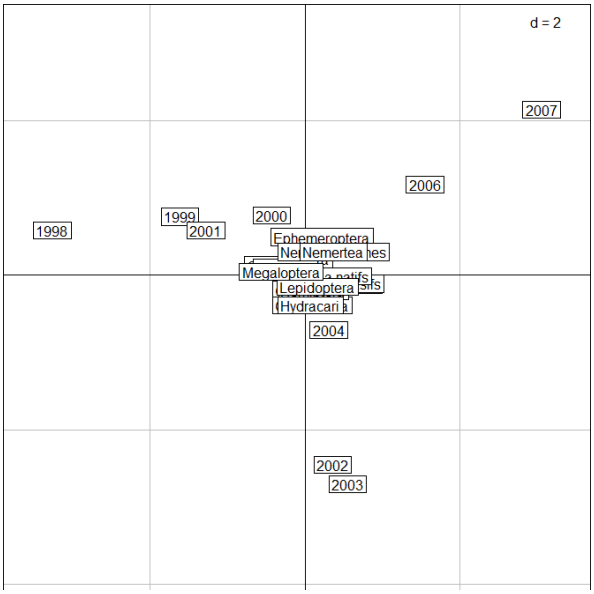


FIGURE 8.20 – Co-inertie entre les macroinvertébrés et les dates

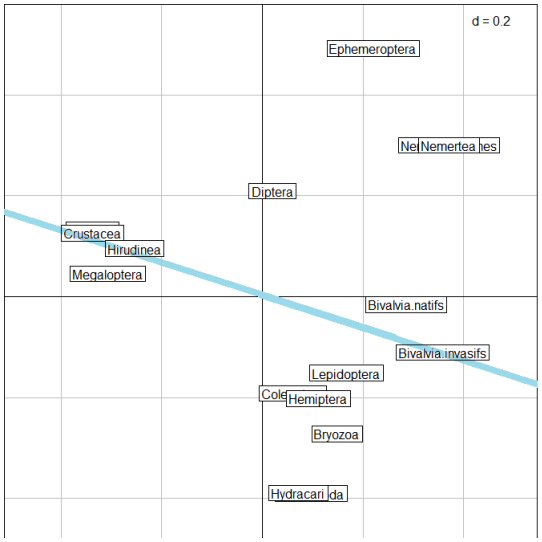


FIGURE 8.21 – Répartition des macroinvertébrés sur le plan

La figure 8.22 présente un résumé de la co-inertie.

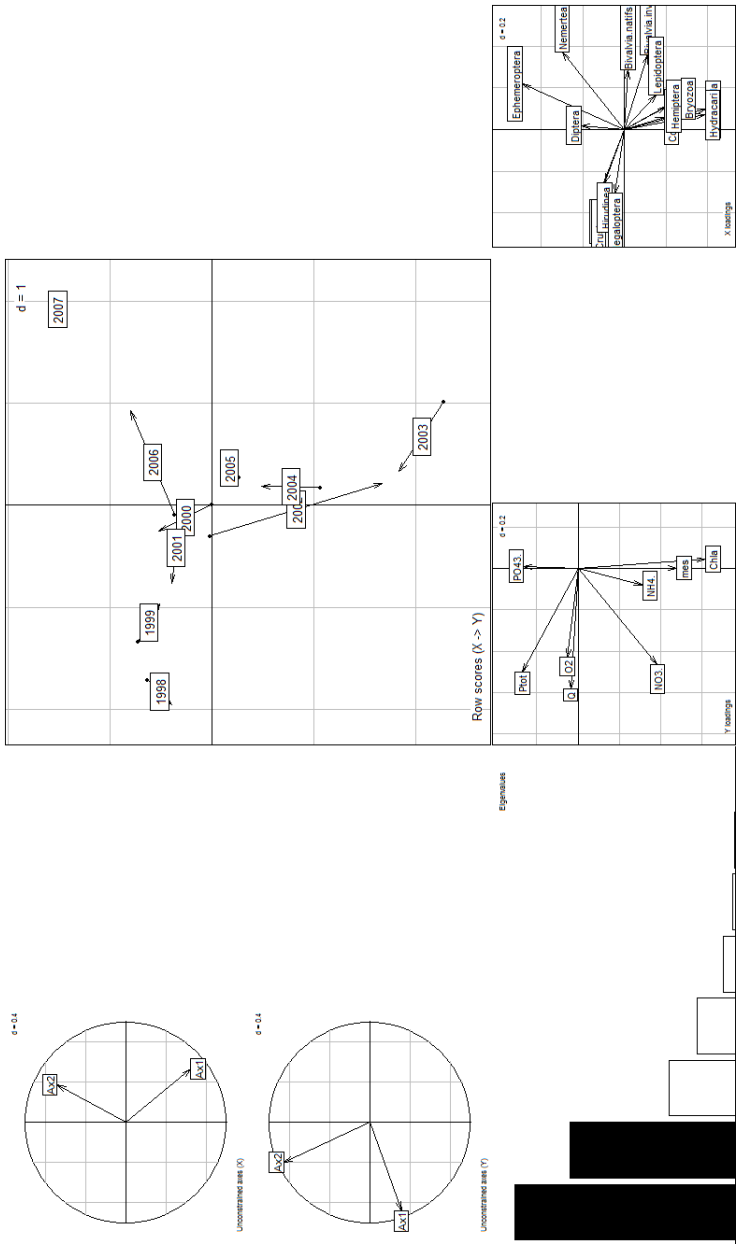


FIGURE 8.22 – Résultats de la co-inertie entre les paramètres physicochimiques et les macroinvertébrés à Sassey

4 Code R

Le code utilisé se situe en annexe 3. Les packages `ade4` et `adegraphics` sont chargés, il s'agit des packages relatifs à la réalisation d'analyses en composantes principales, à l'analyse de co-inertie et la bca (between class analysis).

Un tableau contenant les données est chargé. Les variables (date, paramètres physicochimiques et taxons de macroinvertébrés) sont en colonne, chaque ligne concerne une année d'observation. Nous séparons ensuite les données : un tableau contient les dates, un second les variables physicochimiques et un dernier les taxons de macroinvertébrés.

Nous appliquons une acp sur les deux derniers tableaux grâce à la fonction `dudi.pca`. Ensuite, nous appliquons une bca entre chaque acp et le tableau contenant les dates via la fonction `bca`. Enfin une coinertie est appliquée sur les deux bca obtenues, pour cela nous utilisons la fonction `coinertia`.

5 Conclusion

La co-inertie est un outil formidable permettant d'analyser l'évolution temporelle de nos données. Elle permet une analyse plus précise que l'analyse de tendances, et informe sur l'état « moyen »des données. En plus de nous informer sur l'évolution temporelle de chaque variable, une comparaison aisée peut être réalisée (comme nous l'avons fait pour les bivalves natifs et invasifs) puisque les variables sous toutes représentées sur le même plan.

Chapitre 9

Conclusion

Dans ce mémoire nous abordons une question de biologie : quel est l'impact de bivalves exogènes, appelés bivalves invasifs dans notre étude, sur la biocénose de la Meuse belge et française ? Pour répondre à cette question plusieurs analyses statistiques ont été réalisées. Nous avons étudié deux tableaux de données : d'une part la population de macroinvertébrés regroupée en différents taxons, d'autre part, les mesures de différents paramètres physicochimiques. Les variables les plus importantes sont la chlorophylle a, qui est un bon indicateur de la biomasse de phytoplancton, base de la chaine alimentaire en milieu aquatique ; ainsi que le taxon des bivalves natifs, reprenant les espèces de bivalves autochtones et le taxons des bivalves invasifs, reprenant les espèces de bivalves exogènes, qui sont les taxons ayant une influence directe sur le phytoplancton.

La première analyse réalisée est une analyse de tendance. Cette analyse prend compte des autocorrélations pouvant exister entre les variables. Elle démontre une tendance décroissante significative de la chlorophylle a sur les stations les plus en aval de la Meuse. Le taux de matière en suspension, dont fait partie le phytoplancton, décroit de manière significative sur toutes les stations étudiées sauf celle le plus en amont (Saint-Mihiel). La diminution de chlorophylle a est donc bien réelle.

Une analyse de classification est ensuite effectuée afin de détecter des périodes relatives à la population de bivalves invasifs dans les données. Nous retrouvons deux périodes sur chaque station : la première période, avant 2002, désigne une période où pas ou peu de bivalves invasifs sont détectés, la deuxième période, après 2002, désigne une période où la population de bivalves invasifs est bien installée sur la station. Ainsi l'année 2002 semble être une année charnière.

Afin de mesurer l'impact des bivalves invasifs sur le chlorophylle a, et de comparer cet impact à celui des autres bivalves et des paramètres physicochimiques, des régressions linéaires sont effectuées sur les différentes stations et périodes. Les comportements étudiés sont différents sur chaque station, il est difficile d'en tirer une conclusion générale. En général nous remarquons une influence négative des paramètres physicochimiques sur la chlorophylle a (plus il y a de l'un, moins il y a de l'autre). Nous remarquons également que l'influence des deux types de bivalves est généralement opposée lors de la période après 2002.

La dernière analyse effectuée est une analyse de co-inertie. Celle-ci permet d'appréhender l'évolution des différentes variables étudiées en fonction de l'année de prise des observations. Cette étude confirme les résultats précédents et ouvre d'autres perspectives d'étude plus globale. En effet, une analyse de co-inertie peut être réalisée sur la Meuse entière dont la localisation serait la variable commune, comme nous avons considéré la variable temporelle comme étant la variable commune.

Ce mémoire débouche sur de nombreuses perspectives. En effet, d'autres facteurs peuvent être pris en compte, comme par exemple les facteurs anthropiques ou divers aménagements de la Meuse réalisés sur les différentes stations. Il serait également intéressant de prolonger cette étude à la population piscicole que nous n'avons pas eu le temps d'analyser.

Enfin, nous citerons la publication, actuellement en cours d'écriture, d'Adrien Latli « Drastic

changes in a large European river due to phytoplankton decrease. ». Cette publication reprend des éléments de notre étude, mais elle s'est récemment dirigée vers une étude plus globale sur la Meuse entière tandis que notre étude a une portée plus locale, par station.

Annexe A

Vocabulaire

Anthropique : relatif à l'activité humaine, élément provoqué directement ou indirectement par l'action de l'être humain.
Biocénose : ensemble des êtres vivants qui coexistent dans un même espace.
Bivalve : mollusques d'eau douce dont le corps est couvert d'un coquille.
Consommateur primaire : consommateurs se nourrissant « *d'autres petits animaux, mais ils ne consomment que des herbivores.* »[6].
Consommateur secondaire : « *ensemble organismes capables de synthétiser de la matière organique à partir de matières minérales grâce à l'énergie lumineuse.* »[2].
Exogène (espèce) : espèce qui provient d'un autre pays.
Guilde trophique : ensemble d'espèces ayant les mêmes habitudes alimentaires.
Réseau trophique : « *Ensemble des relations alimentaires entre espèces au sein d'une communauté et par lesquelles l'énergie et la matière circulent* »[3].
Phytoplancton : plancton végétal, ensemble des organismes végétaux vivant en suspension dans l'eau.
Plancton : ensemble des organismes évoluant en suspension dans l'eau [4].
Potamoplancton : plancton vivant dans des courants lents.
Ranq taxonomique : niveau de classification biologie présentée sur la figure 2.1
Taxon : groupe d'organismes vivants.
Turbidité : « *caractère d'une eau dont la transparence est limitée par la présence de matières solides en suspension* »[3].
Trophique : tout ce qui concerne l'alimentation.
Zooplancton : plancton animal.

Annexe B

Annexes concernant les test de tendance

1 Chlorophylle a

TABLE B.1 – Courbe lowess pour les différents taux de chlorophylle a sur les différentes stations

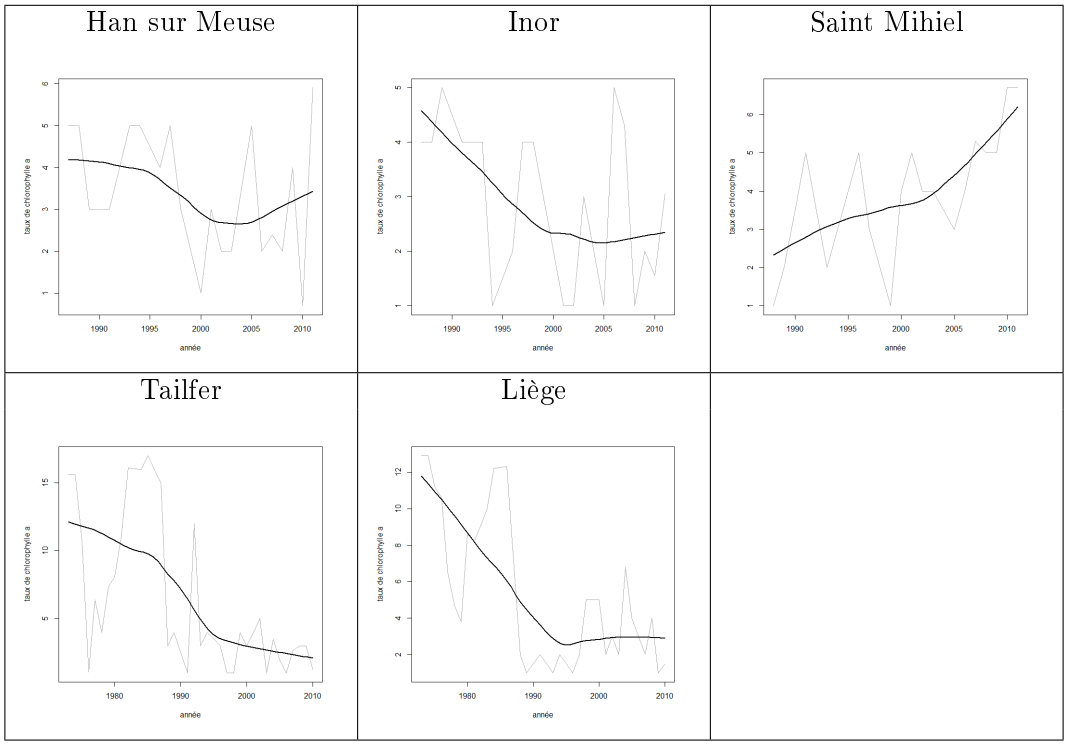
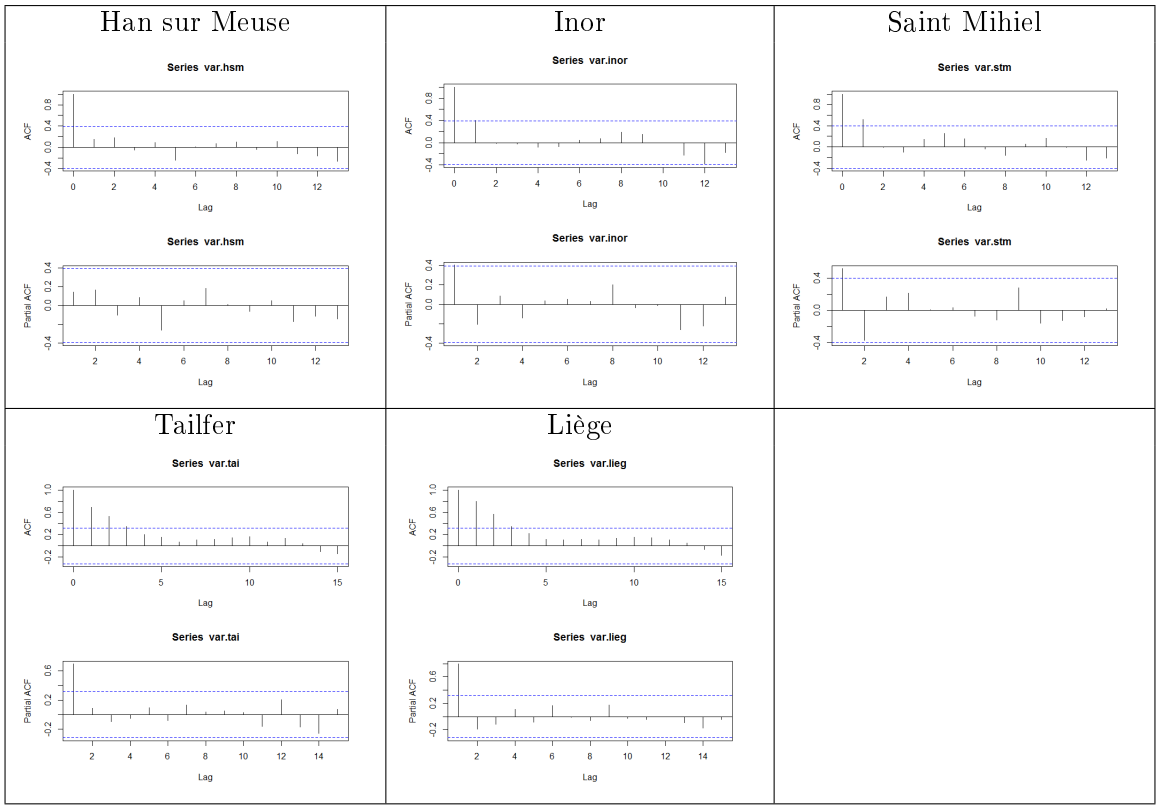


TABLE B.2 – ACF et ACF partielle pour les différents taux de chlorophylle a sur les différentes stations



2 Amonium

TABLE B.3 – Courbe lowess pour les différents taux d'amonium sur les différentes stations

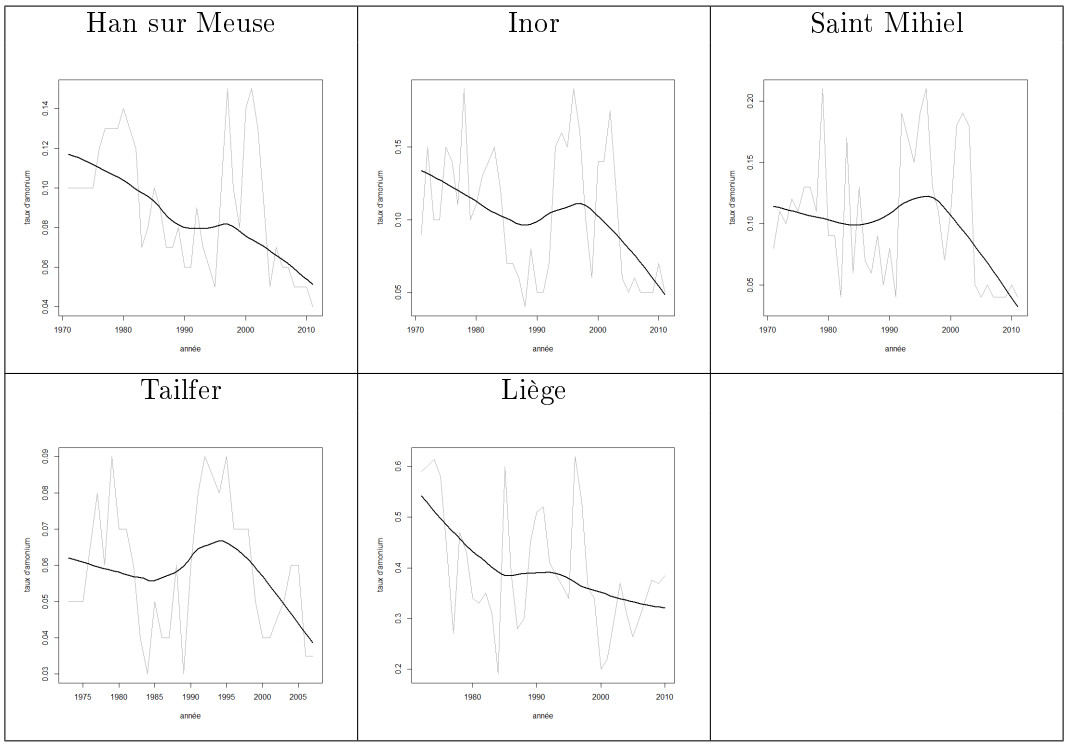
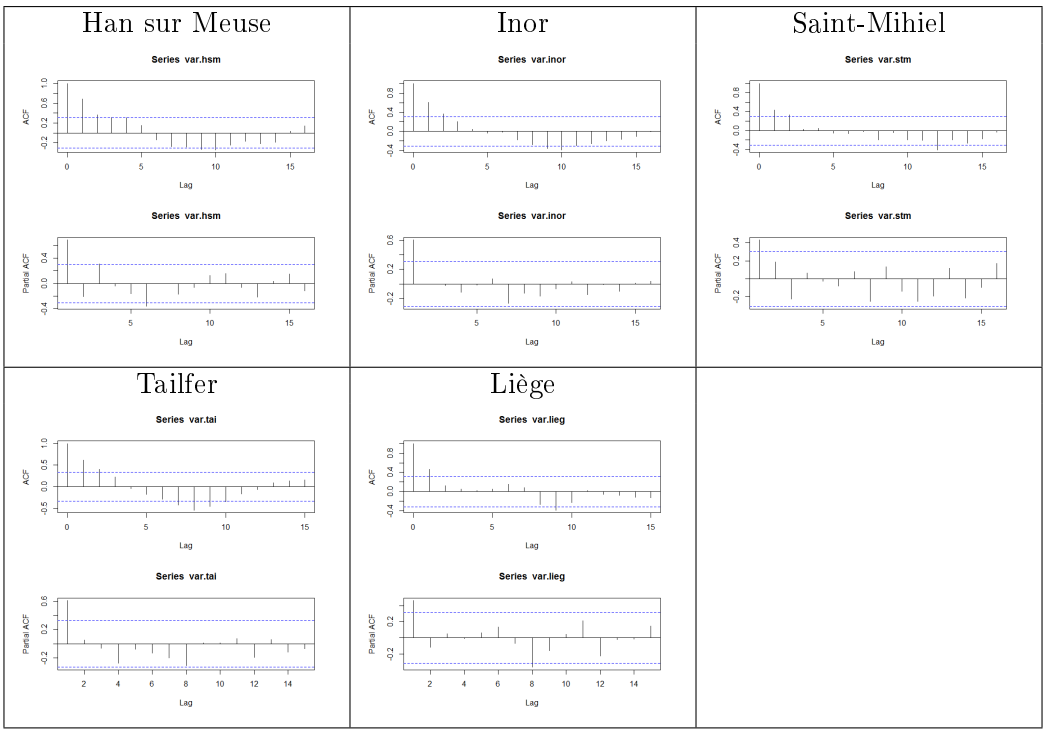


TABLE B.4 – ACF et ACF partielle pour les différents taux d'amonium sur les différentes stations



3 Nitrates

TABLE B.5 – Courbe lowess pour les différents taux de nitrates sur les différentes stations

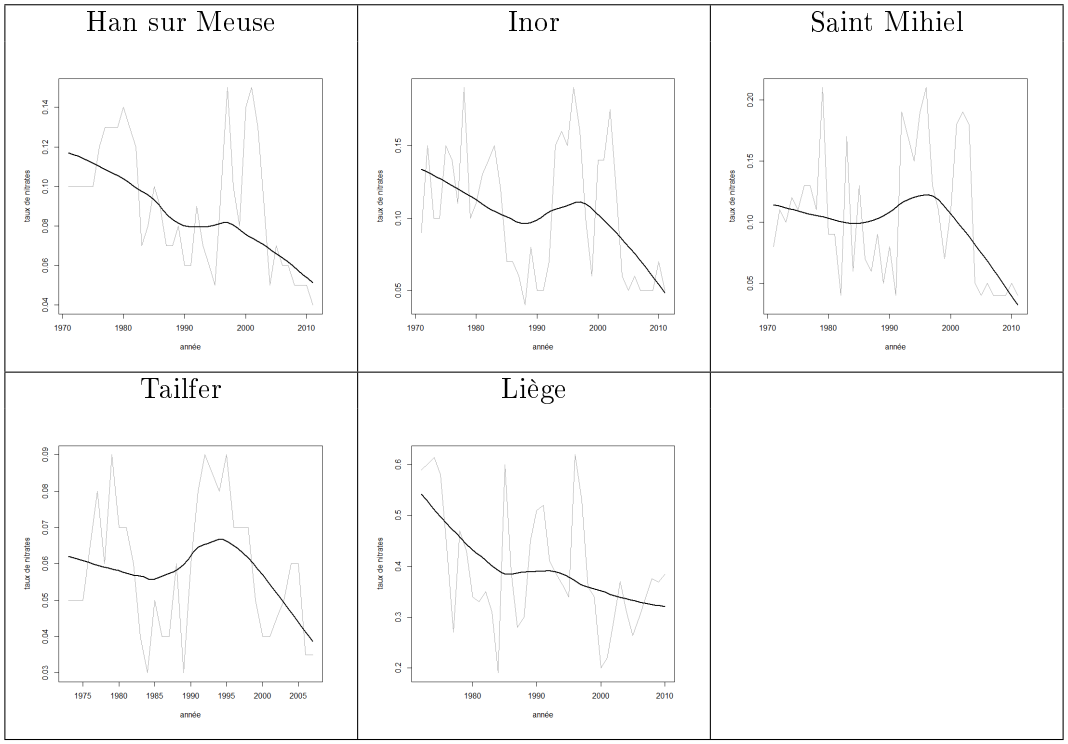
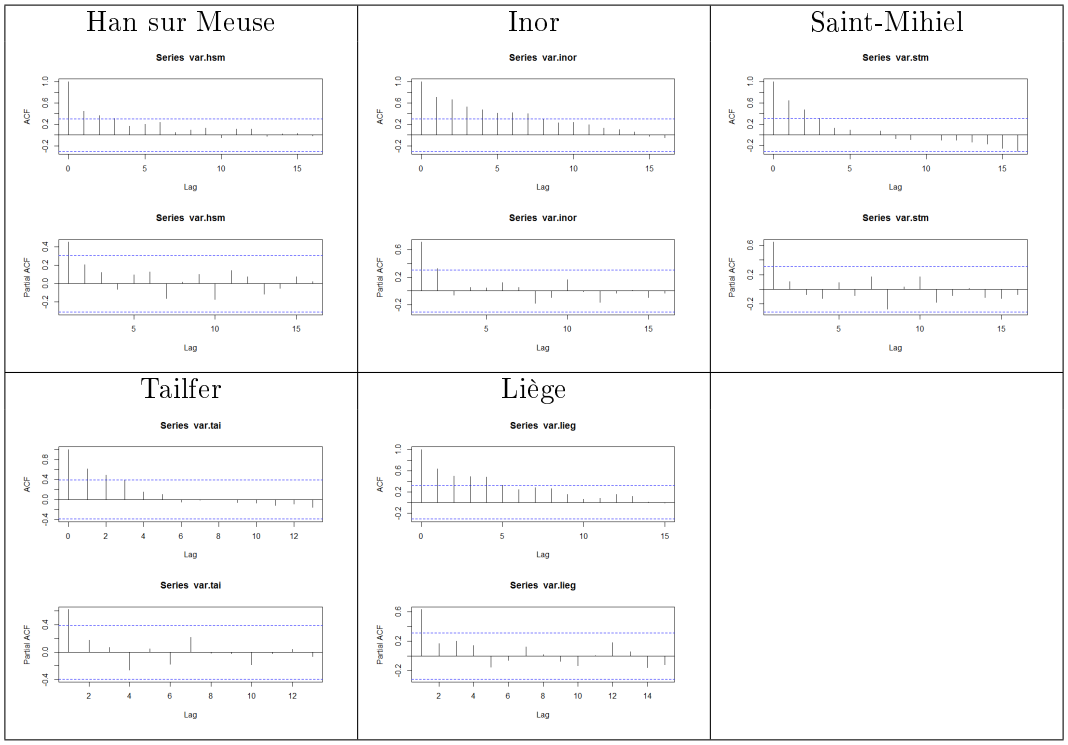


TABLE B.6 – ACF et ACF partielle pour les différents taux de nitrates sur les différentes stations



4 Phosphate

TABLE B.7 – Courbe lowess pour les différents taux de phosphate sur les différentes stations

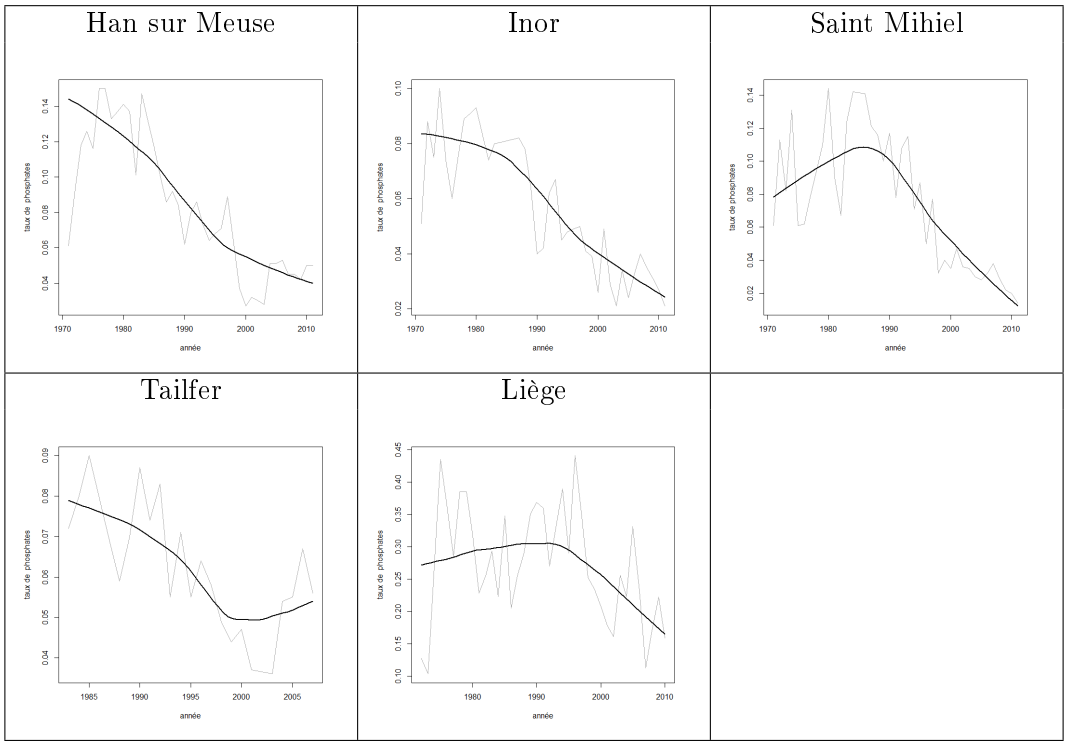
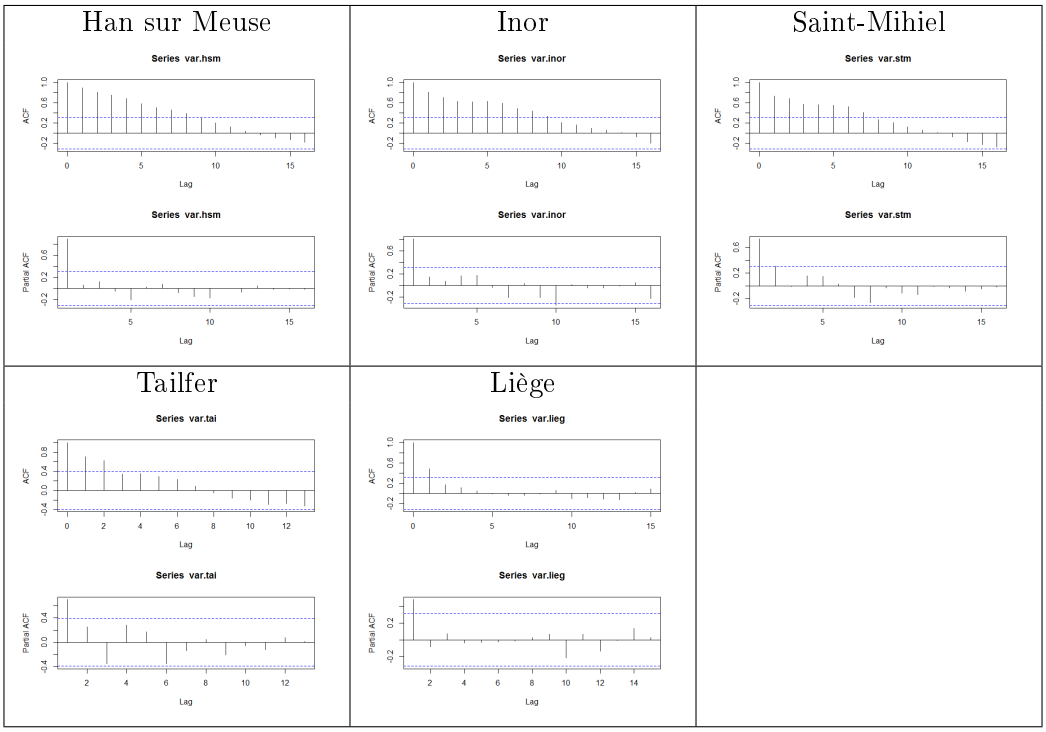


TABLE B.8 – ACF et ACF partielle pour les différents taux de phosphate sur les différentes stations



5 Phosphore

TABLE B.9 – Courbe lowess pour les différents taux de phosphate sur les différentes stations

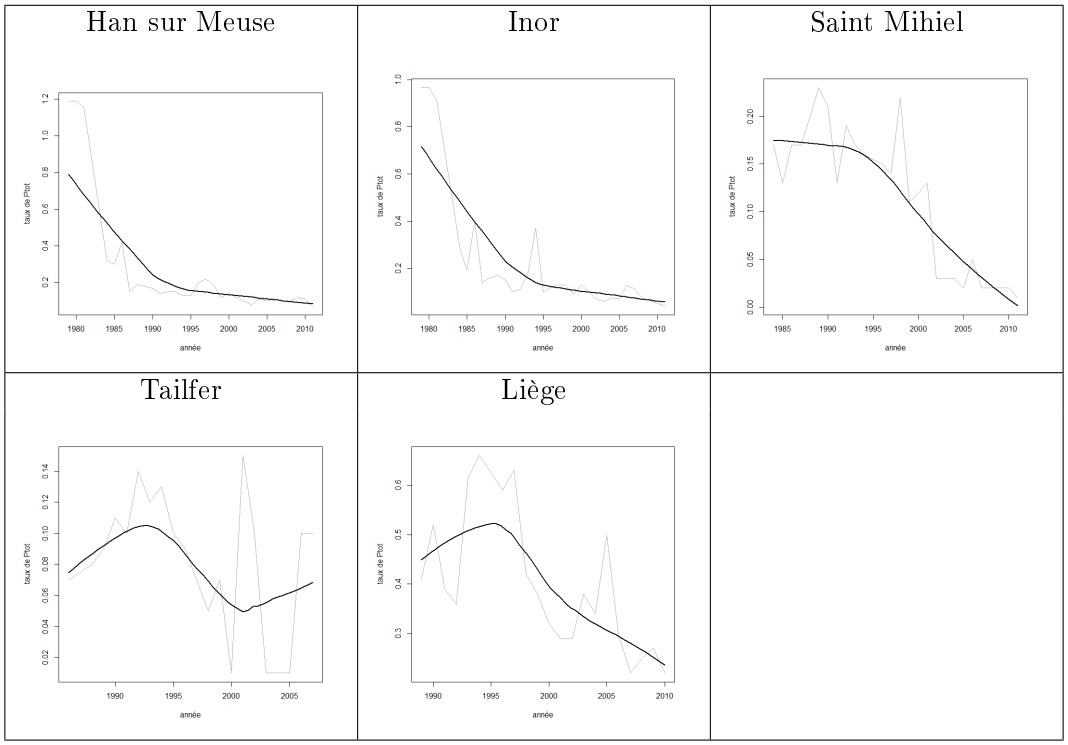
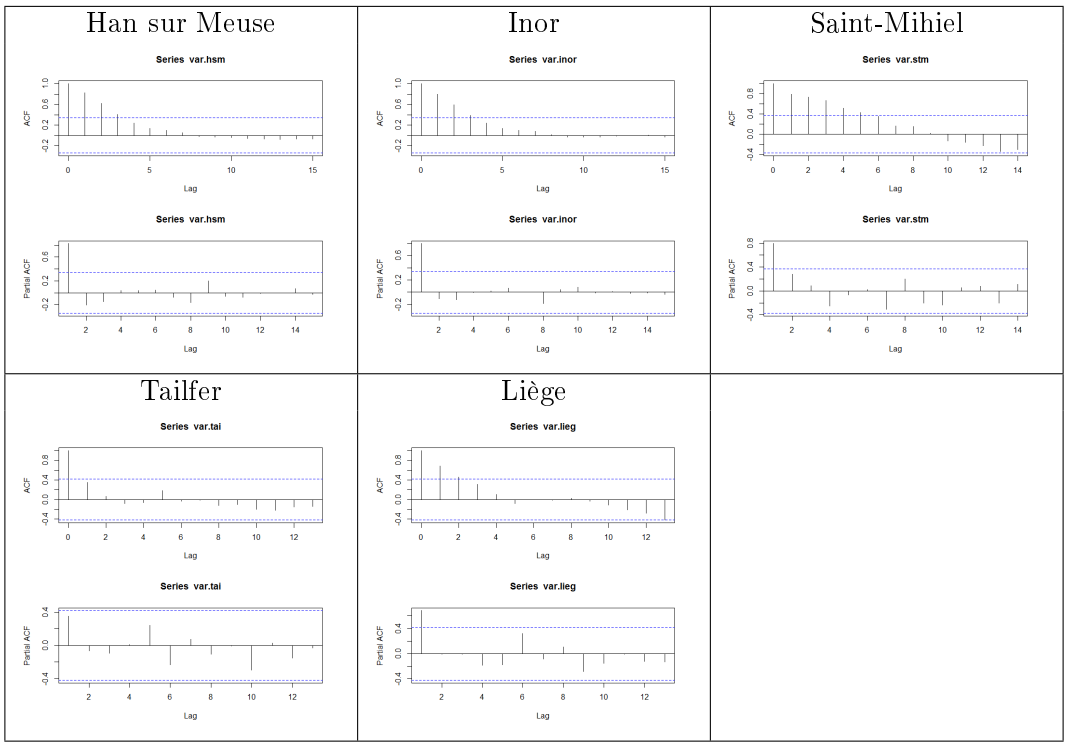


TABLE B.10 – ACF et ACF partielle pour les différents taux de phosphate sur les différentes stations



6 Oxygène

TABLE B.11 – Courbe lowess pour les différents taux d'oxygène sur les différentes stations

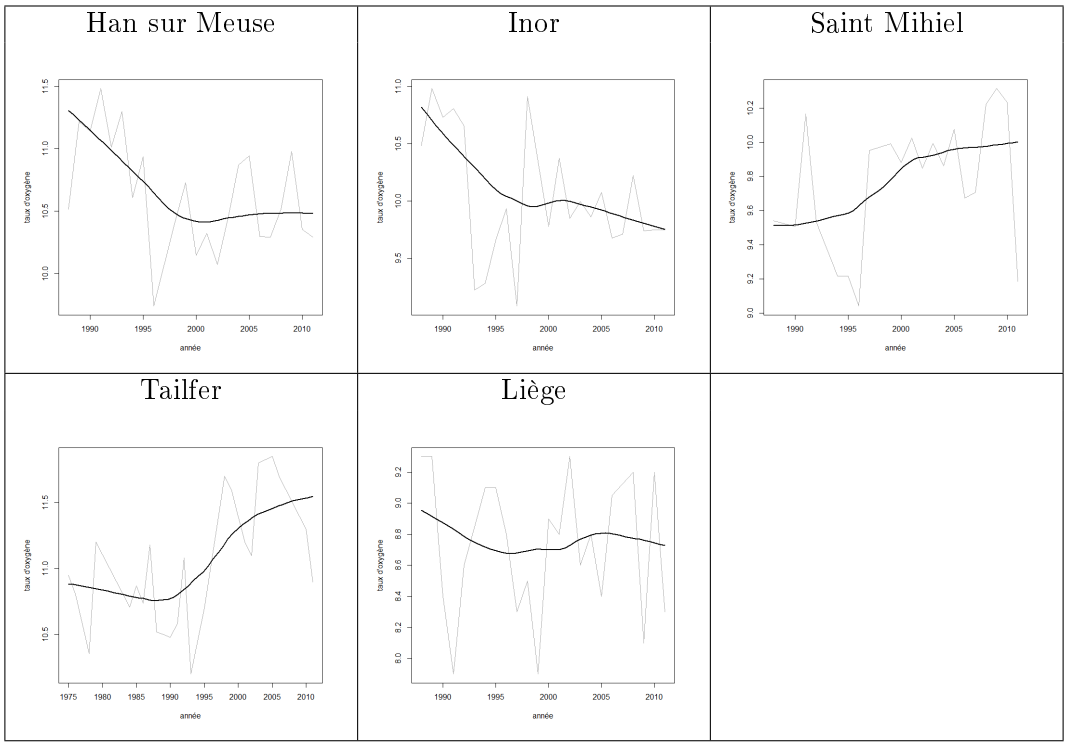
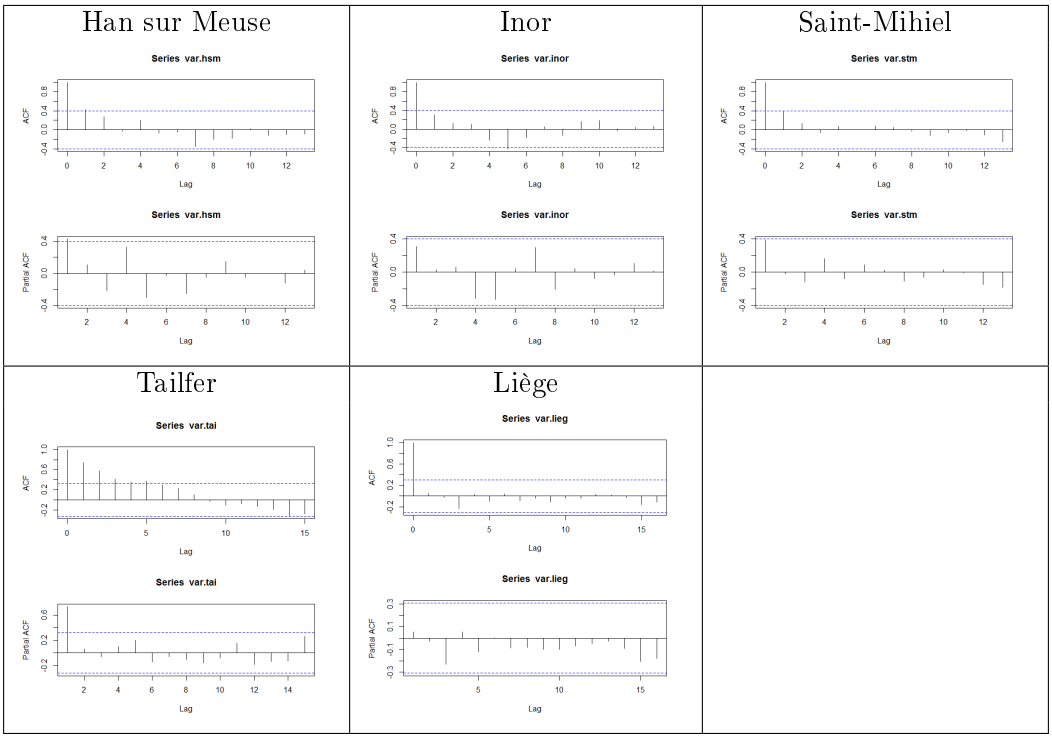


TABLE B.12 – ACF et ACF partielle pour les différents taux d'oxygène sur les différentes stations



7 Q

TABLE B.13 – Courbe lowess pour les différents taux de Q sur les différentes stations

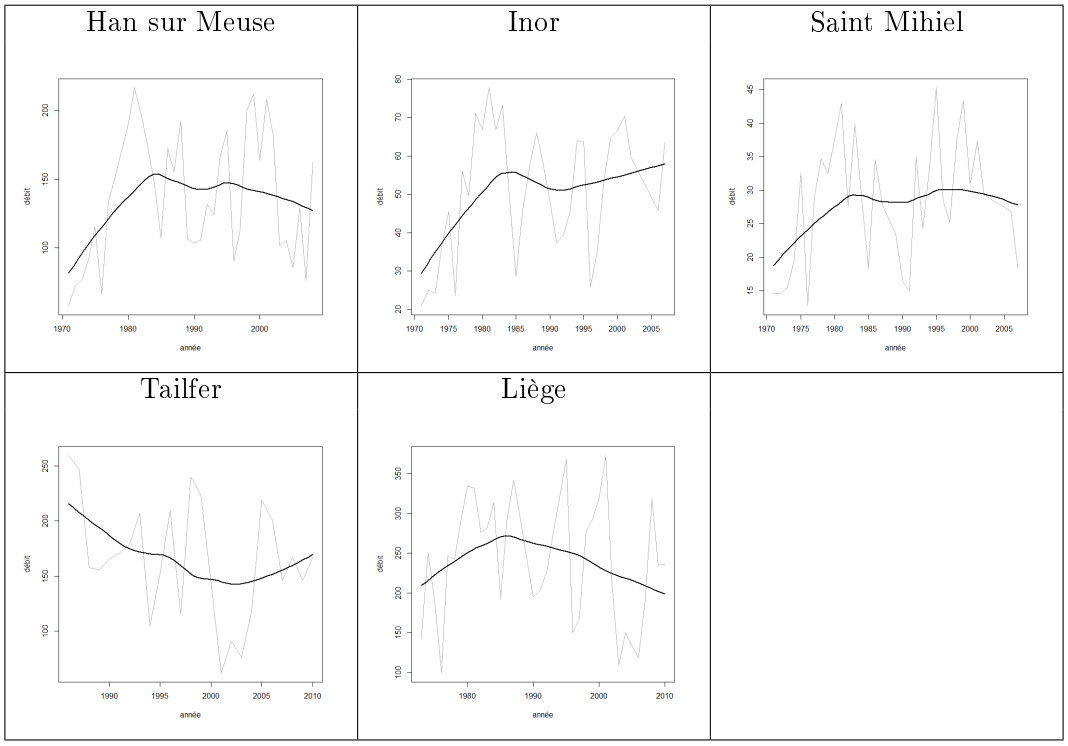
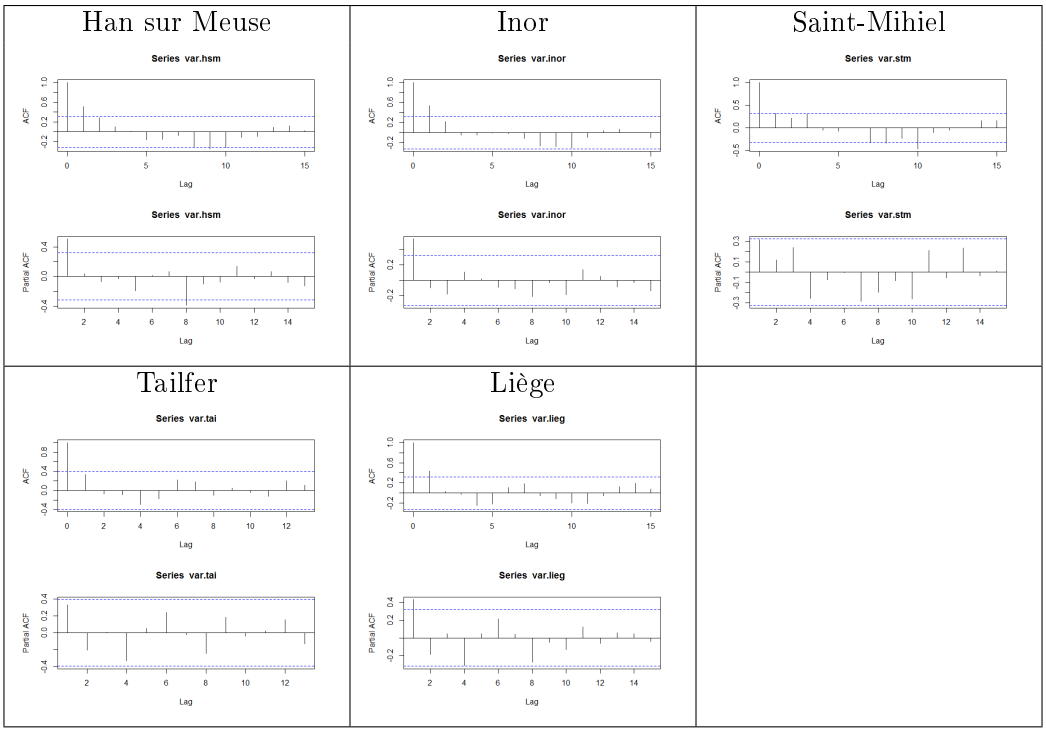


TABLE B.14 – ACF et ACF partielle pour les différents taux de Q sur les différentes stations



8 Température

TABLE B.15 – Courbe lowess pour les différentes températures sur les différentes stations

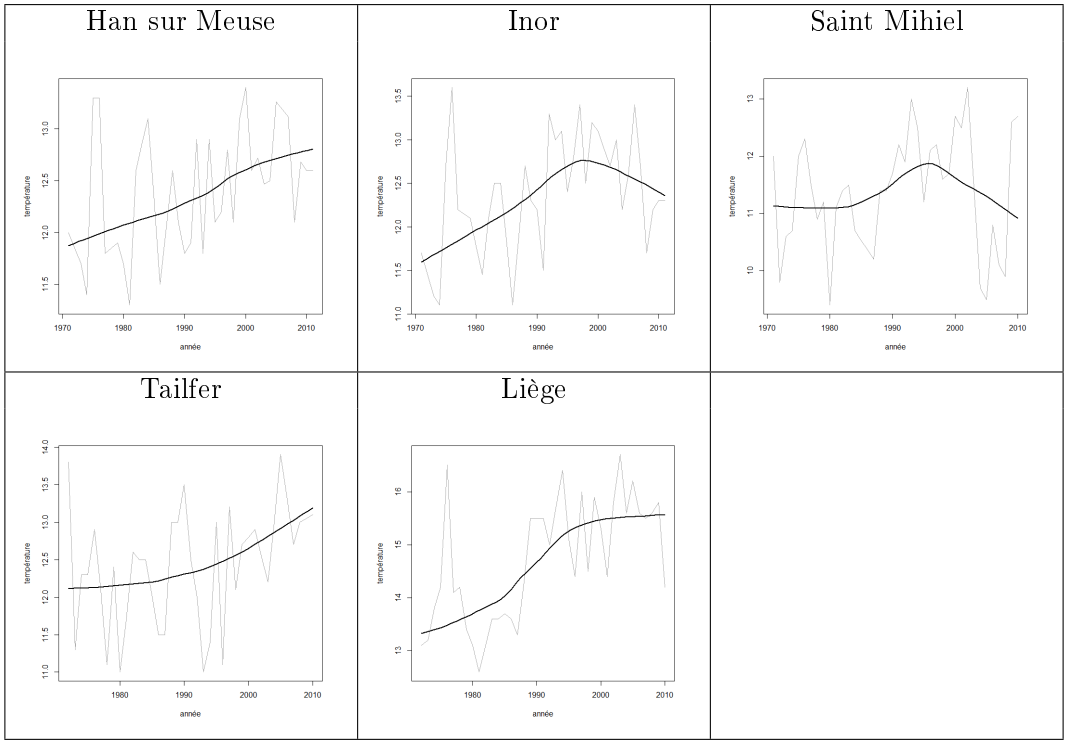
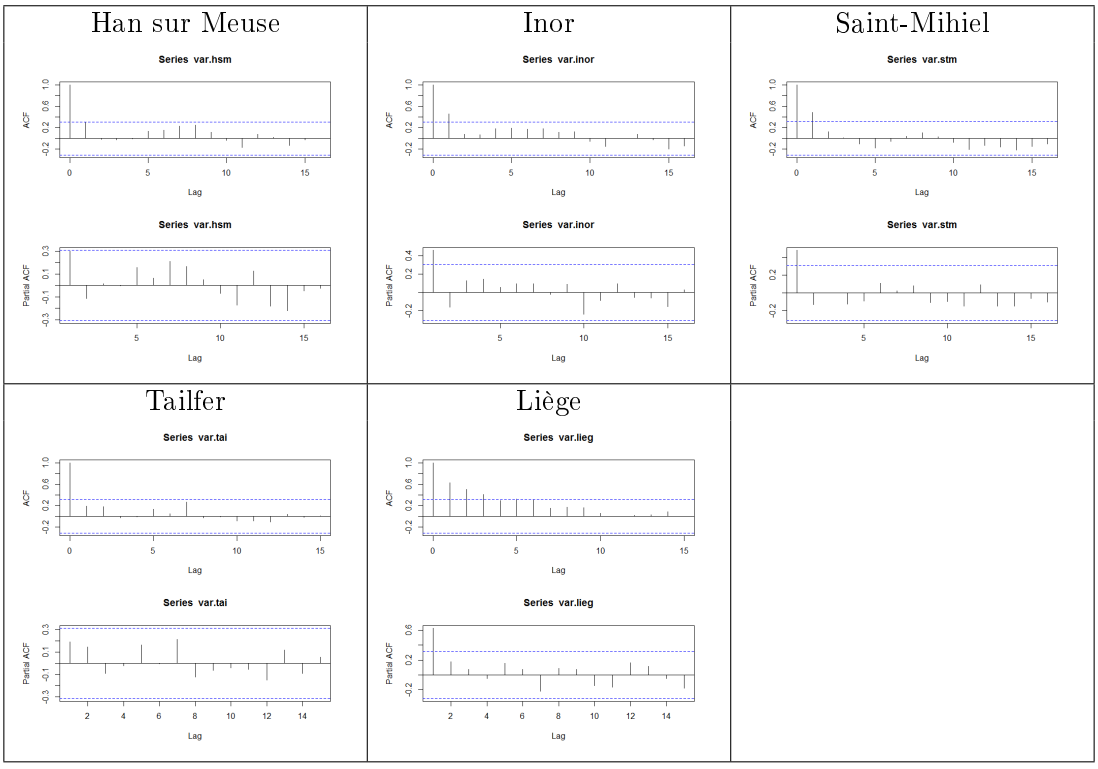


TABLE B.16 – ACF et ACF partielle pour les différentes températures sur les différentes stations



Annexe C

Codes R

1 Détection des tendances par le test de Mann Kendall modifié par Hamed et Rao

```
library(lattice)
library(graphics)
library(Kendall)

#chargement variable
chla=read.table(file="C:/Users/Carol-Ann/Desktop/memoire_R/hamed_rao/chla.csv", header = TRUE
mes=read.table('mes.csv',header=TRUE,sep=';',dec=',',quote="")
nh4=read.table('nh4+.csv',header=TRUE,sep=';',dec=',',quote = "")
no3=read.table('NO3-.csv',header=TRUE,sep=';',dec=',',quote = "")
po43=read.table('PO43-.csv',header=TRUE,sep=';',dec=',',quote = "")
ptot=read.table('Ptot.csv',header=TRUE,sep=';',dec=',',quote = "")
o2=read.table('Oxy.csv',header=TRUE,sep=';',dec=',',quote = "")
q=read.table('Q.csv',header=TRUE,sep=';',dec=',',quote = "")
t=read.table('temp.csv',header=TRUE,sep=';',dec=',',quote = "")




#choix de la variable sur laquelle on travaille (changer le nom du paramètre)
var<-chla

#mettre chaque station dans une variable
var.date=var[1:39,1]
var.stm=var[1:39,2]
var.inor=var[1:39,3]
var.hsm=var[1:39,4]
var.tai=var[1:39,5]
var.lieg=var[1:39,6]

#------------------------------------------------------------------------
#-------------------------- stm -----------------------------------------
#------------------------------------------------------------------------

print('----- stm -----')

newvar.stm<- na.omit(var.stm)  #ne prendre en compte que les variables sans NA
```

```
#calculer la stat. S et tau de Kendall :
out<-Kendall(var.stm,var.date)
summary(out)
score<-out$S       #sauvegarde de la stat. S de MK
varS<-out$varS   #sauvegarde de la variance de S

# calcul des autocorrélations
autocorr<-acf(newvar.stm)        #rend un vecteur
n<-length(autocorr$acf)       #longueur du vecteur rendu
vecteuracf<-autocorr$acf[2:n] #sauvegarde des autocorr dans le vecteuracf

dev.new()
par(mfrow=c(2,1))
acf(var.stm,na.action = na.pass)
pacf(var.stm,na.action = na.pass)

nbredonnees<-length(newvar.stm)   #nbre de var présentes sans compter les NA
sum<-0
print(sum)

for (i in 1:(n-1)){
sum<-sum+(nbredonnees-1)*(nbredonnees-i-1)*(nbredonnees-i-2)*vecteuracf[i]}

cor<-1+2/(nbredonnees*(nbredonnees-1)*(nbredonnees-2))*sum

newvarS<-varS*cor

if(score==0){Znew=0}
if(score>0){Znew=(score-1)/sqrt(newvarS)}
if(score<0){Znew=(score+1)/sqrt(newvarS)}

print('---nouveau Z calculé stm---')
print(Znew)

#-------------------------------------------------------------------------
#--------------------------- inor ----------------------------------------
#-------------------------------------------------------------------------

print('----- inor -----')

newvar.inor<- na.omit(var.inor)  #ne prendre en compte que les variables sans NA

#calculer la stat. S et tau de Kendall :
out<-Kendall(var.inor,var.date)
summary(out)
score<-out$S       #sauvegarde de la stat. S de MK
varS<-out$varS   #sauvegarde de la variance de S

# calcul des autocorrélations
autocorr<-acf(newvar.inor)        #rend un vecteur
n<-length(autocorr$acf)       #longueur du vecteur rendu
vecteuracf<-autocorr$acf[2:n] #sauvegarde des autocorr dans le vecteuracf
```

```
dev.new()
par(mfrow=c(2,1))
acf(var.inor,na.action = na.pass)
pacf(var.inor,na.action = na.pass)

nbredonnees<-length(newvar.inor)   #nbre de var présentes sans compter les NA
sum<-0
print(sum)

for (i in 1:(n-1)){
sum<-sum+(nbredonnees-1)*(nbredonnees-i-1)*(nbredonnees-i-2)*vecteuracf[i]}

cor<-1+2/(nbredonnees*(nbredonnees-1)*(nbredonnees-2))*sum

newvarS<-varS*cor

if(score==0){Znew=0}
if(score>0){Znew=(score-1)/sqrt(newvarS)}
if(score<0){Znew=(score+1)/sqrt(newvarS)}

print('---nouveau Z calculé inor---')
print(Znew)


#------------------------------------------------------------------------
#--------------------------- hsm ----------------------------------------
#------------------------------------------------------------------------

print('----- hsm -----')

newvar.hsm<- na.omit(var.hsm)  #ne prendre en compte que les variables sans NA
n<-length(var.hsm)

#calculer la stat. S et tau de Kendall :
out<-Kendall(var.hsm,var.date)
summary(out)
score<-out$S       #sauvegarde de la stat. S de MK
varS<-out$varS   #sauvegarde de la variance de S

# calcul des autocorrélations
autocorr<-acf(newvar.hsm)        #rend un vecteur
n<-length(autocorr$acf)       #longueur du vecteur rendu
vecteuracf<-autocorr$acf[2:n] #sauvegarde des autocorr dans le vecteuracf

dev.new()
par(mfrow=c(2,1))
acf(var.hsm,na.action = na.pass)
pacf(var.hsm,na.action = na.pass)

nbredonnees<-length(newvar.hsm)   #nbre de var présentes sans compter les NA
sum<-0
print(sum)
```

```
for (i in 1:(n-1)){
sum<-sum+(nbredonnees-1)*(nbredonnees-i-1)*(nbredonnees-i-2)*vecteuracf[i]}

cor<-1+2/(nbredonnees*(nbredonnees-1)*(nbredonnees-2))*sum

newvarS<-varS*cor

if(score==0){Znew=0}
if(score>0){Znew=(score-1)/sqrt(newvarS)}
if(score<0){Znew=(score+1)/sqrt(newvarS)}

print('---nouveau Z calculé hsm---')
print(Znew)



#------------------------------------------------------------------------
#-------------------------- tai -----------------------------------------
#------------------------------------------------------------------------

print('----- tai -----')

newvar.tai<- na.omit(var.tai)  #ne prendre en compte que les variables sans NA

#calculer la stat. S et tau de Kendall :
out<-Kendall(var.tai,var.date)
summary(out)
score<-out$S       #sauvegarde de la stat. S de MK
varS<-out$varS   #sauvegarde de la variance de S

# calcul des autocorrélations
autocorr<-acf(newvar.tai)        #rend un vecteur
n<-length(autocorr$acf)       #longueur du vecteur rendu
vecteuracf<-autocorr$acf[2:n] #sauvegarde des autocorr dans le vecteuracf

dev.new()
par(mfrow=c(2,1))
acf(var.tai,na.action = na.pass)
pacf(var.tai,na.action = na.pass)

nbredonnees<-length(newvar.tai)   #nbre de var présentes sans compter les NA
sum<-0
print(sum)

for (i in 1:(n-1)){
sum<-sum+(nbredonnees-1)*(nbredonnees-i-1)*(nbredonnees-i-2)*vecteuracf[i]}

cor<-1+2/(nbredonnees*(nbredonnees-1)*(nbredonnees-2))*sum

newvarS<-varS*cor

if(score==0){Znew=0}
```

```
if(score>0){Znew=(score-1)/sqrt(newvarS)}
if(score<0){Znew=(score+1)/sqrt(newvarS)}

print('---nouveau Z calculé tai---')
print(Znew)


#------------------------------------------------------------------------
#--------------------------- lieg ---------------------------------------
#------------------------------------------------------------------------

print('----- liege -----')

newvar.lieg<- na.omit(var.lieg)  #ne prendre en compte que les variables sans NA

#calculer la stat. S et tau de Kendall :
out<-Kendall(var.lieg,var.date)
summary(out)
score<-out$S       #sauvegarde de la stat. S de MK
varS<-out$varS   #sauvegarde de la variance de S

# calcul des autocorrélations
autocorr<-acf(newvar.lieg)        #rend un vecteur
n<-length(autocorr$acf)       #longueur du vecteur rendu
vecteuracf<-autocorr$acf[2:n] #sauvegarde des autocorr dans le vecteuracf

dev.new()
par(mfrow=c(2,1))
acf(var.lieg,na.action = na.pass)
pacf(var.lieg,na.action = na.pass)

nbredonnees<-length(newvar.lieg)   #nbre de var présentes sans compter les NA
sum<-0
print(sum)

for (i in 1:(n-1)){
sum<-sum+(nbredonnees-1)*(nbredonnees-i-1)*(nbredonnees-i-2)*vecteuracf[i]}

cor<-1+2/(nbredonnees*(nbredonnees-1)*(nbredonnees-2))*sum

newvarS<-varS*cor

if(score==0){Znew=0}
if(score>0){Znew=(score-1)/sqrt(newvarS)}
if(score<0){Znew=(score+1)/sqrt(newvarS)}

print('---nouveau Z calculé lieg---')
print(Znew)
```

2 Détection des tendances par le test de Mann Kendall et bloc bootstrapping

```
library(outliers)
library(Kendall)
library(boot)

#chargement des données (charger un seul pour tout le programme):
mes=read.table('mes.csv',header=TRUE,sep=';',dec='.',quote="")
chla=read.table('chla.csv',header=TRUE,sep=';',dec='.',quote = "")
nh4=read.table('nh4+.csv',header=TRUE,sep=';',dec='.',quote = "")
no3=read.table('NO3-.csv',header=TRUE,sep=';',dec='.',quote = "")
po43=read.table('PO43-.csv',header=TRUE,sep=';',dec='.',quote = "")
ptot=read.table('Ptot.csv',header=TRUE,sep=';',dec='.',quote = "")
oxy=read.table('Oxy.csv',header=TRUE,sep=';',dec='.',quote = "")
q=read.table('Q.csv',header=TRUE,sep=';',dec='.',quote = "")
t=read.table('temp.csv',header=TRUE,sep=';',dec='.',quote = "")

#choix variable
var<-no3

#mettre chaque station dans un vecteur :

var.stm=var[1,2:42]
var.inor=var[2,2:42]
var.hsm=var[3,2:42]
var.tai=var[4,2:42]
var.lieg=var[5,2:42]



#-----------------------------------------------------
#Transformer les données en séries temporelles:
#-----------------------------------------------------

var.stm<- ts(var.stm, start=c(1971), end=c(2011), frequency=1)
var.stm=var.stm[1,]

var.inor<- ts(var.inor, start=c(1971), end=c(2011), frequency=1)
var.inor=var.inor[1,]

var.hsm<- ts(var.hsm, start=c(1971), end=c(2011), frequency=1)
var.hsm=var.hsm[1,]

var.tai<- ts(var.tai, start=c(1971), end=c(2011), frequency=1)
var.tai=var.tai[1,]

var.lieg<- ts(var.lieg, start=c(1971), end=c(2011), frequency=1)
var.lieg=var.lieg[1,]
```

```
#-------------------------------------------------------
#-------------------Retrait outliers--------------------
#-------------------------------------------------------

var.stm<- na.omit(var.stm)  #ne prendre en compte que les variables sans NA
n<-length(var.stm)

var<-var.stm
n<-length(var)
pv<-0;

while(pv<0.05){
        out<-outlier(var)
        pv<-grubbs.test(var)$p.value
        if (pv<0.05){
        for (i in 1:n){
        if (var[i]==out){
                if(i==1){var[i]<-var[i+1]}
        else{
                if(i==n){var[i]<-var[i-1]}
                else{var[i]<-(var[i-1]+var[i+1])/2}}
}}}}
var.stm<-var

#-------------------------------------------------------

var.inor<- na.omit(var.inor)  #ne prendre en compte que les variables sans NA
n<-length(var.inor)


var<-var.inor
n<-length(var)
pv<-0;
while(pv<0.05){
        out<-outlier(var)
        pv<-chisq.out.test(var)$p.value
        if (pv<0.05){
        for (i in 1:n){
        if (var[i]==out){
                if(i==1){var[i]<-var[i+1]}
        else{
                if(i==n){var[i]<-var[i-1]}
                else{var[i]<-(var[i-1]+var[i+1])/2}}
}}}}
var.inor<-var

#-------------------------------------------------------

var.hsm<- na.omit(var.hsm)  #ne prendre en compte que les variables sans NA
n<-length(var.hsm)


var<-var.hsm
```

```
n<-length(var)
pv<-0;
while(pv<0.05){
        out<-outlier(var)
        pv<-chisq.out.test(var)$p.value
        if (pv<0.05){
        for (i in 1:n){
        if (var[i]==out){
                if(i==1){var[i]<-var[i+1]}
        else{
                if(i==n){var[i]<-var[i-1]}
                else{var[i]<-(var[i-1]+var[i+1])/2}}
}}}}
var.hsm<-var

#-------------------------------------------------------

var.tai<- na.omit(var.tai)  #ne prendre en compte que les variables sans NA
n<-length(var.tai)

var<-var.tai
n<-length(var)
pv<-0;
while(pv<0.05){
        out<-outlier(var)
        pv<-chisq.out.test(var)$p.value
        if (pv<0.05){
        for (i in 1:n){
        if (var[i]==out){
                if(i==1){var[i]<-var[i+1]}
        else{
                if(i==n){var[i]<-var[i-1]}
                else{var[i]<-(var[i-1]+var[i+1])/2}}
}}}}
var.tai<-var

#-------------------------------------------------------

var.lieg<- na.omit(var.lieg)  #ne prendre en compte que les variables sans NA
n<-length(var.lieg)

var<-var.lieg
n<-length(var)


pv<-0;
while(pv<0.05){
        out<-outlier(var)
        pv<-chisq.out.test(var)$p.value
        if (pv<0.05){
        for (i in 1:n){
        if (var[i]==out){
                if(i==1){var[i]<-var[i+1]}
```

```
        else{
                if(i==n){var[i]<-var[i-1]}
                else{var[i]<-(var[i-1]+var[i+1])/2}}
}}}}
var.lieg<-var

#-------------------------------------------------------
#Regarder si les autocorrélations sont significatives
#-------------------------------------------------------

dev.new()
par(mfrow=c(2,1))
acf(var.stm,na.action = na.pass)
pacf(var.stm,na.action = na.pass)

dev.new()
par(mfrow=c(2,1))
acf(var.inor,na.action = na.pass)
pacf(var.inor,na.action = na.pass)


dev.new()
par(mfrow=c(2,1))
acf(var.hsm,na.action = na.pass)
pacf(var.hsm,na.action = na.pass)

dev.new()
par(mfrow=c(2,1))
acf(var.tai,na.action = na.pass)
pacf(var.tai,na.action = na.pass)


dev.new()
par(mfrow=c(2,1))
acf(var.lieg,na.action = na.pass)
pacf(var.lieg,na.action = na.pass)

#-------------------------------------------------------
#-------- Application du test de MannKendall ---------
#-------------------------------------------------------

print('----hsm----')
summary(MannKendall(var.hsm))

print('----inor----')
summary(MannKendall(var.inor))

print('----stm----')
summary(MannKendall(var.stm))

print('----tai----')
summary(MannKendall(var.tai))
```

```
print('----lieg----')
summary(MannKendall(var.lieg))


#-----------------------------------------------------
#-- Bootstrapping pour les variables autocorrélées ---
#-----------------------------------------------------

print('-----------------------------------------------')
print('-------------Block Bootstrapping---------------')
print('-----------------------------------------------')

print('----hsm----')
MKtau<- function(z) MannKendall(z)$tau
boot.out<-tsboot(var.hsm, MKtau, R=1500, l=3, sim="fixed")
#print(boot.out)
#données des tau générés sont dans boot.out$t
summary(boot.out$t)
boot.ci(boot.out,type="perc")
test<-t.test(boot.out$t)
print(test$p.value)

print('----inor----')
MKtau<- function(z) MannKendall(z)$tau
boot.out<-tsboot(var.inor, MKtau, R=1500, l=2, sim="fixed")
#print(boot.out)
#données des tau générés sont dans boot.out$t
summary(boot.out$t)
boot.ci(boot.out,type="perc")
test<-t.test(boot.out$t)
print(test$p.value)


print('----stm----')
MKtau<- function(z) MannKendall(z)$tau
boot.out<-tsboot(var.stm, MKtau, R=1500, l=3, sim="fixed")
#print(boot.out)
#données des tau générés sont dans boot.out$t
summary(boot.out$t)
boot.ci(boot.out,type="perc")
test<-t.test(boot.out$t)
print(test$p.value)

print('----tai----')
MKtau<- function(z) MannKendall(z)$tau
boot.out<-tsboot(var.tai, MKtau, R=1500, l=3, sim="fixed")
#print(boot.out)
#données des tau générés sont dans boot.out$t
summary(boot.out$t)
boot.ci(boot.out,type="perc")
test<-t.test(boot.out$t)
print(test$p.value)
```

```
print('----lieg----')
MKtau<- function(z) MannKendall(z)$tau
boot.out<-tsboot(var.lieg, MKtau, R=1500, l=3, sim="fixed")
#print(boot.out)
#données des tau générés sont dans boot.out$t
summary(boot.out$t)
boot.ci(boot.out,type="perc")
test<-t.test(boot.out$t)
print(test$p.value)
```

3 Analyse de co-inertie

```
library(lattice)
library(ade4)
library(adegraphics)

#-------------------------------------------------------
#---- chargement des données -------------------------
#--------- une colonne = une variable
#-------------------------------------------------------

liege=read.table('reg_liege.csv',header=TRUE,sep=';',dec='.',quote="")

#-------------------------------------------------------
#---- chaque type de variable est placé dans un tableau
#-------------------------------------------------------

date<-liege[,1]
physico<-liege[,2:9]
macro<-liege[,11:26]

date<-as.factor(date)

#-------------------------------------------------------------
#---- acp sur les macroinvertébrés et param. physicochimiques
#-------------------------------------------------------------

acpmacro <- dudi.pca(macro, scan = F, nf = 4)
acpphysico <- dudi.pca(physico, scan = F, nf = 4)

#- diagrammes en batons de la variance expliquée de chaque composante principale
inertie<-acpmacro$eig/sum(acpmacro$eig)*100
barplot(inertie,ylab="% d'inertie",names.arg=round(inertie,2))
inertie<-acpphysico$eig/sum(acpphysico$eig)*100
barplot(inertie,ylab="% d'inertie",names.arg=round(inertie,2))

#-------------------------------------------------------------
#-- bca appliquée la première composante principale de chaque
#-- acp et sur la variable contenant les dates d'observation
#-------------------------------------------------------------
```

```
monbca<-bca(acpmacro,date,scan=F,nf=2)
monbca2<-bca(acpphysico,date,scan=F,nf=2)

#-------------------------------------------------------------
#-- bapplication de l'analyse de co-inertie
#-------------------------------------------------------------

coi<-coinertia(monbca,monbca2)

#--------------------------------------
# graphe résumé de la co-inertie
plot(coi)
#--------------------------------------
# graphe des dates dans le plan
dev.new()
s.label(coi$lY)
#--------------------------------------
# graphe des macroinvertébrés dans le plan
dev.new()
s.label(coi$c1)
#--------------------------------------
# graphe conjoint des dates et param. physico
dev.new()
s.label(coi$lY)                  # graphe des dates
s.label(coi$l1,add = TRUE)       # graphe des param. physico
#--------------------------------------
# graphe conjoint des dates et macroinvertébrés
dev.new()
s.label(coi$lY)                  # graphe des dates
s.label(coi$c1,add = TRUE)       # graphe des macroinvertébrés
```

Bibliographie

[1] Convention sur la diversité biologique, rio de janeiro. https://www.cbd.int/doc/legal/cbd-fr.pdf, 1992.

[2] Dictionnaire environnement. http://www.dictionnaire-environnement.com/producteur_primaire_ID5084.html, consulté le 13 mai 2014.

[3] Dictionnaire environnement. http://www.dictionnaire-environnement.com/, consulté le 8 avril 2014.

[4] Futura nature. http://www.futura-sciences.com/magazines/nature/infos/dico/d/zoologie-plancton-3831/, consulté le 14 mai 2014.

[5] La corrélation de kendall. http://www.jybaudot.fr/Correlations/kendall.html, consulté le 22 octobre 2014.

[6] Les consommateurs secondaires de la chaîne alimentaire. http://www.teteamodeler.com/ecologie/biologie/vivant/consomateur3.asp, consulté le 13 mai 2014.

[7] Mann-kendall test for monotonic trend. http://vsp.pnnl.gov/help/Vsample/Design_Trend_Mann_Kendall.htm, consulté le 31 octobre 2014.

[8] Modèle évolutif r/k. http://fr.wikipedia.org/wiki/Mod%C3%A8le_%C3%A9volutif_r/K, consulté le 14 mai 2014.

[9] Moule zébrée. http://fr.wikipedia.org/wiki/Moule_z%C3%A9br%C3%A9e, consulté le 20 mai 2014.

[10] Autocorrélation. http://fr.wikipedia.org/wiki/Autocorr%C3%A9lation, consulté le 26 février 2015.

[11] Autocovariance. http://fr.wikipedia.org/wiki/Autocovariance, consulté le 5 avril 2015.

[12] The autoreg procedure. http://www.dms.umontreal.ca/~duchesne/chap8.pdf, consulté le 15 mai 2015.

[13] chisq.out.test. http://www.inside-r.org/packages/cran/outliers/docs/chisq.out.testt, consulté le 11 avril 2015.

[14] Classification. http://iml.univ-mrs.fr/~reboul/ADD4-MAB.pdf, consulté le 30 avril 2015.

[15] Classification. http://iml.univ-mrs.fr/~reboul/ADD4-MAB.pdf, consulté le 30 avril 2015.

[16] Cochran's c test. http://en.wikipedia.org/wiki/Cochran%27s_C_test, consulté le 11 avril 2015.

[17] Dixon's q test. http://en.wikipedia.org/wiki/Dixon%27s_Q_test, consulté le 11 avril 2015.

[18] Durbin–watson statistic. http://en.wikipedia.org/wiki/Durbin%E2%80%93Watson_statistic, consulté le 15 mai 2015.

[19] Grubbs' test for outliers. http://en.wikipedia.org/wiki/Grubbs%27_test_for_outliers, consulté le 11 avril 2015.

[20] Outlier. http://www.mathworks.com/matlabcentral/fileexchange/13439-orbital-mechanics-library/content/Groundtrack.m, consulté le 24 avril 2015.

[21] Outlier. http://en.wikipedia.org/wiki/Outlier, consulté le 26 avril 2015.

[22] Phytoplancton. http://fr.wikipedia.org/wiki/Phytoplancton, consulté le 10 mai 2015.

[23] Processus autorégressif. http://fr.wikipedia.org/wiki/Processus_autor%C3%A9gressif#Moments_d.27un_processus_AR.281.29, consulté le 15 mai 2015.

[24] Processus stochastique. http://fr.wikipedia.org/wiki/Processus_stochastique, consulté le 5 avril 2015.

[25] R-square statistics and other measures of fit. http://support.sas.com/documentation/cdl/en/etsug/60372/HTML/default/viewer.htm#etsug_autoreg_sect023.htm, consulté le 15 mai 2015.

[26] Regression with autocorrelated errors. http://support.sas.com/documentation/cdl/en/etsug/60372/HTML/default/viewer.htm#etsug_autoreg_sect003.htm, consulté le 15 mai 2015.

[27] Régression linéaire. http://fr.wikipedia.org/wiki/R%C3%A9gression_lin%C3%A9aire, consulté le 26 février 2015.

[28] Régression (statistiques). http://fr.wikipedia.org/wiki/R%C3%A9gression_lin%C3%A9aire, consulté le 24 avril 2015.

[29] Série temporelle. http://fr.wikipedia.org/wiki/S%C3%A9rie_temporelle, consulté le 1 mai 2015.

[30] Série temporelle. http://www.onema.fr/IMG/pdf/2011_032.pdf, consulté le 1 mai 2015.

[31] Test de kolmogorov-smirnov. http://fr.wikipedia.org/wiki/Test_de_Kolmogorov-Smirnov, consulté le 15 mai 2015.

[32] Test de shapiro-wilk. http://fr.wikipedia.org/wiki/Test_de_Shapiro-Wilk, consulté le 11 avril 2015.

[33] Valeur p. http://fr.wikipedia.org/wiki/Valeur_p, consulté le 3 avril 2015.

[34] Unamur A. Hardy. *Statistique*. 2012.

[35] D. Chessel and A.-B. Dufour. Co-inertie, co-structures et compromis. http://pbil.univ-lyon1.fr/R/pdf/cssb8.pdf, consulté le 20 mai 2015.

[36] Strayer D.L. and Smith L.C. *Relationships between zebra mussels (Dreissena polymorpha) and unionid clams during the early stages of the zebra mussel invasion of the Hudson River.* Freshwater biology, 1996.

[37] Sheng Yue et Paul Pilon. *Hydrological Sciences Journal : A comparison of the power of the t test, Mann-Kendall and bootstrap tests for trend detection.* Taylor & Francis, 2004.

[38] Darrigran G. *Potential impact of filter-feeding invaders on temperate inland freshwater environments ; Biological invasions vol. 4.* Kluwer academics publisher, 2002.

[39] R.O. Gilbert. *Statistical Methods for Environmental Pollution Monitoring.* van nostrand company Inc., 1987.

[40] A. Hardy. *Cours de Régression linéaire et non linéaire.* Unamur, 2011-2012.

[41] A. Hardy. *Cours de classification.* Unamur, 2013-2014.

[42] Moulthon J. *Répartition du genre Corbicula Megerle von Mühlfeld (Bivalvia : Corbiculidae) en France à l'aube du XXI siècle. Hydrocologie appliquée vol. 12.* 2000.

[43] Cécile Delattre. EDP Sciences. Jeremy Alonso, Sébastien Mougenez. *Tendances d'évolution du peuplement de poissons de la Meuse à Chooz et 1991 à 2008.* 2014.

[44] Pierre Jost. Traitement des points aberrants. http://infochimie.u-strasbg.fr/master/Cours_stat_pdf/PT_ABERR.PDF, consulté le 11 avril 2015.

[45] Descy J.P. and al. *Modélisation du bilan en oxygène dans la Moselle.* Freshwater biology, 1993.

[46] Neha Karmeshu. Trend detection in annual temperature and precipitation using mann kendall test - a case study to asses climate change on select states in the northeastern united states. http://repository.upenn.edu/cgi/viewcontent.cgi?article=1045&context=mes_capstones, consulté le 1er novembre 2014.

[47] Otjacques W. Latli A. and Kestemont P. *Etat des stocks de poissons en Meuse belge, identification des causes de déclin et proposition de mesures de remédiation.* 2000.

[48] J.R. Lobry. Analyse de co-inertie sur données simulées et sur données protéomiques. http://pbil.univ-lyon1.fr/R/pdf/tdr641.pdf, consulté le 20 mai 2015.

[49] Martial Ferreol Cecile Delattre et Yves Souchon Mathieu Floury, Philippe Usseglio-Polatera. *Global climate change in large European rivers : long-term effects on macroinvertebrate communities and potentieal local confounding factors.* Blackwell publishing, 2012.

[50] V. Monbet. Tests statistiques. http://fr.wikipedia.org/wiki/Test_de_Kolmogorov-Smirnov, consulté le 15 mai 2015.

[51] D. Chessel S. Dray. Le couplage de tableaux écologiques. http://pbil.univ-lyon1.fr/Stage2013/Perfectionnement2013/Partie4.pdf, consulté le 21 mai 2015.